THE MORAL STATUS OF WHOLE BRAIN EMULATIONS

PADRAIC XAVIER GIDNEY

A thesis of 16,168 words (including footnotes) submitted in partial fulfilment of the requirements for the degree of Bachelor of Arts (Honours)

INTRODUCTION

Artificial Intelligence is going to radically change our world; the only real question is by how much. A number of prominent figures believe that current AI research might initiate a so-called technological singularity - a period where intelligent machines design even more intelligent machines, setting off an exponentially accelerating cascade of advancement whose end result, a superintelligence, would be "the last invention that man need ever make" (Good 1965). However, even for those who dismiss such singularity talk as hyperbolic sci-fi nonsense, the fact that we're on the cusp of an AI revolution - and that society is going to look very different once it's over - seems undeniable. Already AI systems are changing how we eat[1], how we transport people and goods[2], how we diagnose and treat illnesses[3], and how we wage war[4]. They are replacing and outperforming humans in a plethora of tasks, many of which were once thought to require a uniquely human "instinct"[5], and their scope of application only looks to be increasing.

The response of many thinkers to this apparent reality of an impending, AI-driven social upheaval has been to focus on how to navigate the process safely. There have, for example, been no shortage of public figures[6] and popular articles[7] discussing the problem of making sure that intelligent machines, though autonomous, still end up behaving in ways we'd approve of. A particular question that has dominated the philosophy of artificial intelligence has been the problem of how to get AIs to understand human ethical norms, and - if this were possible - which ethical norms to instil in them[8].

---

[1] Hampton Creek (a silicon valley startup in the US that is now valued at over $1.1 billion) is using machine learning to identify plant substitutes for animal-based products. They've already used this technique to develop the world's first store-stocked vegan mayonnaise, as Bosker (2017) reported.
[2] Google has already invested over $1 billion into their "Waymo" self-driving car project, and Tesla is expected to announce a self-driving electric lorry in late October of this year - facts that were reported by Ohnsman (2017) and Gibbs (2017a) respectively.
[3] As Bloch-Budzier (2016) reported, Google's DeepMind partnered with the UK's National Health Service last year to set up an AI system that would build comprehensive patient profiles from data collected during blood tests and GP visits in order to more efficiently identify anomalies for early treatment.
[4] The United States National Security Agency already uses machine-learning algorithms to identify terrorist "couriers" using communications meta-data (Robbins 2016), and the Department of Defence has relied on artificially intelligent systems to manage their logistics since at least 1991, when the DART system was introduced to help run the Desert Storm operation (Bostrom 2014, p. 19).
[5] Prime examples would be the recent triumph of Google's 'AlphaGo' system over the world's best Go player (Shead 2017) as well as the recent success of the University of Alberta's 'Deepstack' neural network in beating professional poker players at Texas Hold'Em (Caruso 2017) - both games traditionally conceived of as requiring a great deal of intuition.
[6] See, for example, the recent submission of an open letter to the UN signed by Elon Musk and 115 other notables calling for a total ban on autonomous robots being employed in warfare (Gibbs 2017b).
[7] Clear examples include the raft of articles discussing how self-driving cars ought to react in emergencies (e.g. Lin 2013) or how the technological singularity might be avoided or else survived (e.g. Urban 2015; Khatchadourian 2015).
[8]This was, for example, the focus of chapters twelve and thirteen of "Superintelligence" (Bostrom, 2014), as well as sections five and six of "The Singularity: A Philosophical Analysis" (Chalmers, 2010).

A comparatively neglected question, both in popular culture and in the philosophical literature, is whether and to what extent the new class of artificial entities that this revolution will produce might not themselves deserve some moral consideration[9]. Should AIs have rights, for example? Should it matter how we treat them? Should they *count* in our decision-making?

The object of this thesis is to tackle these questions as they pertain to a particular category of artificially intelligent systems called 'whole brain emulations' (henceforth WBEs). A WBE is essentially a very accurate computational model of a living creature's brain which is created by scanning that brain at incredibly fine resolutions (capturing basically every detail down to the neuronal level), and then reproducing the scanned structure in a sophisticated simulation. Such a simulation would, if successful, behave in almost exactly the same way as the original biological brain and so would, to any observer with only indirect access to the system, appear to be an indistinguishable "reproduction of the original intellect, with memory and personality intact" (Bostrom 2014, p. 36). No human WBEs exist yet (current scanning and simulation technologies are simply not up to the task), however - as Nick Bostrom notes (Ibid., p. 36) - there are "no fundamental conceptual or theoretical" obstacles to their creation, and it is effectively only a matter of time before one is developed[10].

The fact that WBEs represent the most inevitable route to an artificial general intelligence or 'AGI' (i.e. an AI system capable of performing every cognitive task a human would be capable of with equal or greater proficiency) is certainly one reason to focus on them in our inquiry, but there are at least two more strong reasons to do so. In the first place, WBEs seem much more likely than other sorts of AI to be conscious (their behaviour and neural blueprint being nearly identical to our own), and it seems *prima facie* plausible that this might be a precondition for having moral standing. These intuitions might be wrong, of course, and would have to be inspected much more closely (as indeed they will be) before any conclusions were drawn, but they do point to WBEs as a good place to start the discussion on the moral status of intelligent systems.

A second reason why focusing on WBEs seems sensible is that they will probably be the kinds of AI that humans interact with most intimately. Whereas other kinds of AGI would appear to be mere tools for solving human problems, WBEs would have actual human personalities and desires (being, as they are, simulations of humans who once existed). It seems highly probable that they would be a part of our community, even if they might not be treated on a par with flesh and blood humans, and so it seems likely that they would come up in our moral decision-making far more often. Naturally, then, we should want to establish our ethical duties towards WBEs before moving on to consider what duties we might also owe to other forms of AI with whom our interactions will almost certainly be less direct and meaningful.

---

[9] In "Machine Ethics" (Anderson & Anderson 2011), which provided a review of the eponymous field, only one of thirty-one chapters related even indirectly to this question.

[10] A complete emulation of the nervous system of the roundworm *Caenorhabditis elegans* is expected in the next five years. David Dalrymple, who is working on this project, stated that he "would be extremely surprised … if this was still an open problem in 2020" (Bostrom 2014, p. 332)

The overall position which this thesis aims to defend with respect to WBEs is that there is no way of determining their true moral status, with the result that we might never be entirely sure how to act around such entities, or what place they ought to occupy in our communities. My argument for this position consists of three parts. In section 1, I argue that the moral status and conscious stats of any entity are intimately linked, so that we cannot know the former without first knowing the latter; In sections 2, I argue that the conscious status of WBEs cannot be determined empirically; and in section 3, I argue that main non-empirical approach for determining their conscious status also fails. Insofar as these arguments are successful, this suffices to establish that WBEs are, for all intents and purposes, moral question marks. I then consider - in the thesis' postscript - whether there might not still be some way of factoring them into our moral decision-making. I consider what I take to be the most natural proposal for doing so, and provide a brief sketch of why this proposal will not suffice.

SECTION I - MORAL STATUS AND CONSCIOUS STATUS

The object of this section is to defend the claim that we cannot determine an entity's moral status until we have first determined its conscious status. The section has three parts: in the first, I explain the particular sense of 'moral status' being employed (moral patienthood, effectively), provide a rough and ready definition of phenomenal consciousness, and sketch the various existing theories about the link between these two concepts. Next, I introduce Carruthers' "Phenumb" thought experiment and outline a modified version of it which, I argue, can serve as a litmus test for any theory that would deny this section's central claim. Finally, I present a series of reasons why those theories must necessarily fail that litmus test, and why the claim must therefore be accepted.

Laying the Conceptual Groundwork

A first point of importance is that when we use the term 'moral status' in this essay, we are really just referring to moral patienthood. Moral patienthood can be intuitively understood as the property of "showing up on a moral radar screen" (Gruen 2017) or, just as well, of being the sort of entity who ought to be treated as "an end in itself, a being with inherent value" (Korsgaard 2007, p. 5) rather than a mere means to the ends of others. A more precise definition, and the one that we will use going forward, is that to be a moral patient is to be a source of normative claims on moral agents and, in particular, to give such agents a reason not to interfere with you.

This definition needs a bit of unpacking. Firstly, it should be noted that the kind of reason which moral patienthood generates applies to all moral agents equally, and independently of any special relationship they might have with the moral patient in question - a child's moral patienthood, for example, would give their parents just as much reason to refrain from hurting them as it would a total stranger[11]. A second important point is that this definition of moral patienthood makes reference to "moral agents", and yet that term also seems to require a definition (and ideally one that does not itself refer to moral patienthood). There are several definitions of moral agents one might choose from[12], but for the purposes of this paper this category will be left somewhat imprecise, except insofar as it should be understood to include - at the very least - ordinary, mentally and emotionally unimpaired adult human beings. The practical upshot of all this is that, when this essay asks whether a particular entity has any moral status, this should be read as asking whether that entity is the sort of being than an ordinary adult human should consider themselves obliged to leave alone, all else equal.

[11] This is not to say that parents ought to treat their children like strangers (an absurd position, surely); it is to say that the additional moral reasons which parents have for caring for their children do not derive from their children's inherent moral status, but from the specific child-parent relationship the two share.
[12] One intuitive proposal from Christine Korsgaard is that a moral agent is any entity which faces "the normative question" (Korsgaard 1992, p. 23) - i.e. which can consider its reasons for acting and ask itself if they are good ones.

A second matter to be dealt with before our actual arguments begin is the question of what precisely we mean by the term phenomenal consciousness. This is a fair enough question, but a complete answer to it will not be given here primarily because, as Ned Block notes, it seems impossible to give a 'precise' definition of phenomenal consciousness "in any remotely non-circular way" (Block 1995, p. 230). Indeed, when it comes to consciousness, the best way we have of getting clearer about the notion seems to be simply pointing at different instances of it[13]. To this end, let us consider a series of expressions, metaphors, and examples that - taken together - might serve as a sort of a rough and ready definition of phenomenal consciousness. The expressions that have been used most often and most helpfully to try and capture this phenomenon include Thomas Nagel's claim that all it is for an entity to be conscious is for there to be "something it is like" (Nagel 1974, p. 436) to be that entity; Peter Carruthers' idea that conscious states are those which have a "subjective feel" about them (Carruthers 2005, p. 84); and Robert Gullick's suggestion that consciousness is something with a unique "qualitative character" or "raw feel" to it (Gullick 2017). Of course, ordinary talk about consciousness rarely makes use of these academic phrases. When non-philosophers try and define what consciousness is, they tend instead to lean on a series of metaphors which include, for example, the idea of 'the lights being on' inside conscious entities, or the idea of there 'being someone in there'. As a last resort, it can be useful to point to actual instances of consciousness and its absence. This is the approach Searle takes when he explains that conscious experiences are "those subjective states of awareness or sentience that begin when one wakes in the morning and continue … until one falls into a dreamless sleep, into a coma, or dies"[14] (Searle 1990, p. 635). Together, this motley collection of gestures towards phenomenal consciousness will serve as our operating definition of the phenomenon.

Having now provided a sketch of the concepts of moral status and phenomenal consciousness as they shall be understood in this thesis, we can turn to the various theories which concern the relationship between them. These can be divided into three broad categories:

I. Firstly, there are theories which support a strong connection between consciousness and moral standing. In this group would be those that tie moral standing directly to a capacity for consciousness (e.g. the sentience based accounts of Korsgaard (2004; 2007) and Singer (1993)); theories which tie it to this capacity ***plus*** some other quality (e.g. Tom Regan's "experiencing subjects of a life" account (Regan 1985, p. 24), and Michael Tye's meta-cognitive account of suffering (Tye 2000, p. 182)); and theories which support such features as central, but which extend moral standing to entities which have only an indirect relationship to them (e.g. entities who merely have the potential to be conscious, or who were once conscious, or who are part of a biological kind that is conscious).

[13] Block notes this as well, even going so far as to say that "all one can do is point" (1995, p. 230).
[14] An almost identical definition is given by the prominent neuroscientific researcher of consciousness Giulio Tononi, who remarks that consciousness "is what vanishes every night when we fall into dreamless sleep and reappears when we wake up or when we dream" (Tononi 2008, p. 216).

II. Secondly, there are theories which are ambiguous about the connection between consciousness and moral standing. In general, these link moral standing to a particular feature which is itself ambiguously related to consciousness. A key example of this sort of view would be accounts on which moral standing is grounded in autonomy (e.g. Kant 1775, Quinn 1984), since it is unclear whether a being would really count as willing things autonomously, or willing things at all, if it were unconscious. Other examples would be accounts that ground moral standing in a capacity for self-awareness (Tooley 1972), in having a 'well-being' (Quinn 1984, again), or in being part of a social community (Anderson 2004), each of which might well be available or unavailable to non-conscious entities depending on our specific conception of the property in question. The best approach to thinking about theories like this (and the approach that will be adopted going forward) is to treat each theory as having two versions: a conscious version, on which the central proposed property is deemed to involve or require consciousness, and a non-conscious version, on which the property and consciousness are interpreted as being independent. This effectively dissolves this category into the two others.

III. Lastly, there are theories which explicitly deny any connection between consciousness and moral standing. Chief among these are theories which ground moral standing in the ability to form preferences about things (a view that is very close to the non-conscious version of a Kantian autonomy theory, and which finds support in Peter Carruthers (1999; 2004)), as well as certain theories in environmental ethics[15] which claim that moral standing is essentially a matter of being naturally occurring and un-created (Elliot 1997), so that everything from the microorganisms to plants to human beings (but obviously not artificial intelligences like WBEs) will have moral standing.

The Phenumb Thought Experiment: A Litmus Test

We now turn to the central argument of this section, which is that we should (contra the 'no connection' theories of moral standing, as well as the non-conscious readings of the 'unclear connection' theories) endorse a strong connection between consciousness and moral standing. More precisely, I argue that we should assent to the following claim: that if two entities differ only in the following way - that one is conscious, while the other has never been, never will be, and is not part of a kind that is conscious - then the moral status of the conscious entity will be significantly higher, so that interfering with it in the same way will in general be ethically worse. It follows as a corollary to this claim that we will not be able to determine the moral status of an entity until we can determine its conscious status[16].

[15] It may still be, on such theories, that it's worse to interfere with a conscious being than a non-conscious one, but this will only be because there are certain ways we can interfere with conscious entities (e.g. by causing them pain or frustration) that we can't interfere with non-conscious ones (e.g. plants). In principle, if the interference was the same in both cases, both entities would have an equal claim against it.

[16] Where conscious status is to be construed as meaning, not just whether the entity is conscious, but also - as above - whether that entity ever has been, could be, or is member to a kind that is conscious.

In making this case, I'm going to focus chiefly on the 'Phenumb' thought experiment that Carruthers raises, developing it to show that any theory which denies a strong connection between consciousness and moral standing will end up producing highly unintuitive results. Before introducing that thought experiment, however, it behooves us to go into slightly more detail concerning Carruthers' view of the relationship (or lack thereof) between consciousness and moral standing. His position is that interfering with a philosophical zombie by frustrating their desires (i.e. the outcomes they would choose to pursue), or causing them non-conscious 'equivalents' of pain or other negative emotions (i.e. states with the same neural and cognitive profile as our pain and negative emotion, but with no associated phenomenology) would be essentially as unethical as interfering in the same way with a regular, conscious human[17]. The underlying intuition motivating this view is that what makes these sorts of interference ethically important (i.e. deserving of sympathy) is not how they cause the interfered-with entity to feel, but that they cause it to believe that some state of affairs to which it is strongly averse has come to pass.

It was to buttress this intuition that Carruthers first developed the Phenumb thought experiment (Carruthers 1999, p. 478). The experiment revolves around a hypothetical entity called 'Phenumb', who is very like a normal human, except does not feel any conscious desire-satisfaction or desire-frustration (though he does still have other conscious experiences, e.g. of vision and sensory pain). When Phenumb achieves a goal, he does not experience the warm glow of exultation that we do, but he will report something like 'that was a thing I greatly desired, and worthwhile to have obtained', and likewise in failure. The question which Carruthers asks us to consider is how should we treat Phenumb and his various projects. His answer is worth quoting in full (Ibid, p. 479):

"When Phenumb has been struggling to achieve a goal and fails, it seems appropriate to feel sympathy: not for what he now feels - since by hypothesis he feels nothing, or nothing relevant to sympathy - but rather for the intentional state which he now occupies, of dissatisfied desire … Similarly, when Phenumb is engaged in some project which he cannot complete alone, and begs our help, it seems appropriate that we should feel some impulse to assist him"

The conclusion which Carruthers ultimately arrives at is that "the psychological harmfulness of desire-frustration has nothing (or not much…) to do with phenomonology, and everything (or almost everything) to do with thwarted agency" (Ibid., p. 478). Moreover, he notes that this clearly has nothing to do with the fact that Phenumb could feel other things consciously, e.g. physical pain, since these other conscious experiences played no essential role in the example. Hence, Carruthers thinks, both philosophical zombies and non-human animals (which - for quite separate reasons - he thinks are non-conscious) are appropriate objects of sympathy and concern, at least insofar as they form preferences.

[17] What he says, specifically, is that the question he is interested in is "whether those inventions of the philosophical imagination, zombies, would be appropriate objects of sympathy and concern" (Carruthers 1999, p. 467) - the answer he arrives at being a definite 'yes'.

In his 2004 paper, Carruthers extends this line of reasoning to cover pain and negative emotions too. He argues that such states all involve both a non-conscious perceptual component[18] as well as a felt component, and that the 'awfulness' and hence ethical seriousness of these states stems - even in ordinary humans - from the former. When we are stressed, for example, Carruthers thinks this involves, in the first place, the perception of a cluster of unwelcome physical symptoms such as tight-chestedness and sweating (which he collectively refers to as the "bodily gestalt" (2004, p. 116) of stress) and, in the second place, the unique phenomenological feeling of stress. According to him, it is the first of these that makes stress so unpleasant, and in virtue of which it is morally wrong to inflict stress upon others. His arguments here are quite weak[19] and the position itself does not seem hugely plausible, since it seems intuitively obvious that the reason we are so averse to states like stress or pain is precisely because of the way they *feel*. This need not spell defeat for Carruthers, however, because all he really needs for his overall position to succeed is that a non-conscious entity ***could*** find their non-conscious versions of pain and other negative emotions (i.e. their perceptions of the associated bodily gestalts) as aversive as we find ours. If this weaker claim were true - and it doesn't seem obviously false - then, in combination with his claim that aversiveness is what is really ethically important (the claim his 1999 Phenumb case was designed to convince us of), we would get the desired result that zombies deserve just as much sympathy as their conscious counterparts.

I find the Phenumb thought experiment remarkably illuminating, and what I propose to do now is to expand upon it slightly to build it into what is, in effect, a litmus test for all the theories of moral standing which deny a strong link between that phenomenon and consciousness. To do this, let us imagine a creature called 'Phenumb+' (henceforth P+) who, like Phenumb, has conscious visual, tactile, auditory, etc. experiences while lacking any feelings of desire satisfaction or dissatisfaction, but - unlike Phenumb - is incapable of having any valent conscious experiences at all (including the negative quality of pain or emotions like sadness). P+ is, in other words, a 'fully blanched' version of Phenumb. When P+ strikes his finger with a hammer, he perceives the same "bodily gestalt" that we do (e.g. the feeling of a large amount of pressure being applied, and physiological symptoms like throbbing), but he won't experience the familiar phenomenal 'feeling' of pain. Likewise, when he is in a stressful situation, he may perceive that he has become tight-chested and sweaty, but will not experience the phenomenology of stress that we experience above and beyond these physical symptoms. Now, for the sake of argument, let us further suppose that P+ nonetheless finds his non-conscious versions of pain and stress (and whatever other negative emotions there are) just as aversive as we find out conscious version.

---

[18] Essentially this involves perceiving all the physical symptoms that ordinarily go along with pain and other negative emotions. For example, Carruthers suggests (2004, p. 117) the perceptual component of stress might involve noticing that one is sweaty and tight-chested.

[19] They trade on an equivocation of 'aversive' (in the preference-satisfaction sense) and 'awful' (in the ordinary sense), as well as on the assumption that Carruthers' widely disbelieved 'Higher Order Thought' theory of consciousness is correct.

The reason that P+, so described, can act as a litmus test for those theories which deny a link between that moral standing and consciousness, is that each of those theories - in order to deny such a link - assert that moral standing is grounded in some specific capacity or property (e.g. the capacity to will, or to be aware of oneself as a self, or to have a wellbeing, etc.) that is itself supposed to be independent of consciousness. Observe, though, that P+ possesses each of these non-conscious capacities and properties to just the same degree that an average human does, and so must - if any of those accounts is correct - have exactly the same moral standing as an ordinary person. What this means, in effect, is that every one of the denialist theories of moral standing must hold that interfering with P+ in some way is just as ethically problematic as interfering with an normal person in the same way.

A Strong Connection Between Consciousness and Moral Standing

Now that the groundwork has been laid, and the entity P+ fleshed out, I will argue that it is actually absurd to hold that interfering with P+ as ethically problematic as interfering in the same way with an ordinary human, with the result that each theory which denied a strong link between consciousness and moral standing must be false.

A first point to note is that P+ will not behave in exactly the same way as a normal human[20] precisely because some normal human behaviour is a result of the felt quality of our experiences which P+'s experiences lack. Consider, for example, how P+ would react to his non-conscious version of pain. As was mentioned above, he would still find this experience (i.e the perception of pain's "bodily gestalt") highly aversive, and so would still be inclined to take the same sort of deliberate, thought-through steps as a normal person would to alleviate it (e.g. he might put a burn under water, or bandage a wound). P+ will also react to painful stimuli with essentially the same sort of reflexes as humans do (e.g. the quick pulling away of digits and limbs), since these are largely controlled by the spinal cord and brainstem - anatomical features which P+ shares with us, since they have nothing to do with generating valent experiences. There are, however, a plethora of pain-related behaviours that are non-reflexive, and which seem to be a direct result of the 'feeling' of pain rather than the mere desire to be rid of it (which is clear since many of these actions are in fact counterproductive, inhibiting efficient pain-relief). Examples of such behaviours include rocking back and forth, doubling over, groaning, yelling 'ouch', and rubbing a sore area[21]. It seems fairly clear that, since such behaviours are a result of the phenomenology of pain, and since P+'s pain lacks this phenomenology, these behaviours will not be a normal part of his pain response (though he might, of course, voluntarily decide to exhibit them if that suited his purposes).

[20] This is one of the many ways in which P+ differs from a traditional philosophical zombie. The other important difference is that his neurology would lack those mechanisms which, in us, are responsible for generating valent experiences.

[21] Such behaviours are, when exhibited by animals, taken as strong evidence (even definitive evidence, on some views, though I will question this idea in the following section) that the animal is experiencing conscious pain rather than merely unconscious nociception. For example many people point to Sneddon's 2003 paper - which showed that fish, injected in the lip with a noxious liquid, rub their lip on the gravel and rock back and forth on the spot - as proof that fish feel conscious pain.

With this in mind, let us construct a new thought experiment. Imagine how P+ would react if he were kidnapped by a sadist and tortured mercilessly. He would obviously find the whole affair, as well as the various analog bodily properties that he'd perceive once the torture began (the bodily gestalts associated with pain and distress), highly undesirable, which means he'd be motivated to take any action which he believed would improve his circumstances. We can suppose, for the sake of argument, that the actions open to him are few (each limb is bound) and that his torturer - being, as he is, a sadist - would only be encouraged by outward displays of pain (e.g. shouting, wriggling, gritting of teeth) so P+ would refrain from such acts, at least so far as he is able. Importantly, and unlike ordinary humans, P+ actually is able to refrain from the lion's share of such behaviours, since - as was just mentioned - these behaviours are, by and large, the product of a painful phenomonology which is entirely absent in him. True, there will be a few localized reflexes controlled by the spinal cord (e.g. pulling back briefly when a new painful stimulus is applied), but largely P+ will remain entirely still throughout - totally unflapped, by all outward appearances. What's more, we know this isn't just an appearance; P+ actually is unbothered by the situation, at least in the sense that he doesn't feel either positively or negatively about it (rather than in the strangely intellectual sense of 'bothered' as simply being disposed to take action that would remove him from the situation). Importantly, P+ is not like some hardened special forces soldier who hardly winces through the ordeal, but who we know is - on the inside - in a world of phenomenal suffering. On the contrary, P+'s occasional jerks and twitches, far from being a window into a tumultuous inner experience, are mere mechanistic responses - more at odds with, than reflective of, his mind's contents (chief among which is his desire to remain as still as possible). If P+ had it his way, he'd remain still as a plank until the police arrived, then hop off the torture-bench and carry on his merry way, this undesirable episode happily (or rather, desirably) over.

In considering the above thought-experiment, the question we should be asking is not 'should we sympathize with P+?', since we could still give an affirmative answer to that question even if we thought that P+ lacked moral standing entirely (this would be the case, for example, if we felt that beholding any form of humanoid suffering with indifference cultivated a callousness of character which would then be damaging in one's relations with conscious humans[22]). We should also not be asking whether P+'s pain deserves sympathy in its own right (the question Carruthers puts to us), since, although interesting, it has no bearing on the matter of a strong connection between consciousness and moral standing[23]. The question we should really ask is how our sympathy for P+ should compare to our sympathy towards a human put through the same ordeal. Imagine, for example, that the aforementioned torturer had kidnapped two

[22] This is a worry attributable to Kant, who believed that animals lacked moral standing. On his view, malicious action directed towards animals, while not intrinsically wrong, "gradually uproots a natural disposition that is very serviceable to morality in one's relations with other people" (in Korsgaard 2004, p. 90)

[23] Suppose we know that P+'s pain does matter. This only tells us that P+ has moral standing, but it might still be that he has less moral standing than he would if he could experience this pain consciously, which is the point in contention.

individuals, P+ and some regular human passerby, and tormented both of them - the one (P+) lying impassively for the experience to be over, the other (the conscious human) writhing about in agony. Suppose, further, that we could free one of these prisoners. Who would we pick? The answer is obvious, of course, and takes us no time at all - we free the human, whose pain is not just intellectual, but felt! To put a finer point on it, we might return to the single-kidnapping case (where only P+ is being tortured) and ask ourselves whether we would, for a dollar, flip a switch that suddenly rendered P+ phenomenally conscious (by, say, rewiring his brain). The monetary amount is irrelevant, so long as it remains small, because the point is clear: experiencing pain phenomenally is no trifling ethical matter.

Of course, it is open to Carruthers - at this stage - to abandon his claim as it relates to pain and negative emotions, and retreat to the weaker claim that frustrating P+'s desires is, all else equal, just as ethically problematic as frustrating the desires of a conscious human (i.e. the claim he first made in his 1999 paper). To defeat this claim, however, we need only look to the philosophical literature around happiness, and in particular to Fred Feldman's hypothetical involving 'Glum' the depressed philosophy graduate. Feldman (2010, p. 65) asks us to imagine a person whose numerous important desires are all believed, by them, to be failing to materialize. In actuality, this person is faring much better - with respect to those desires - than they give themselves credit for, but from their own perspective everything is falling apart. Following therapy, Glum comes to form "a much more realistic view of his talents" (Ibid., 65) and, as a result, to believe that his desires are being satisfied to a far greater degree. However, despite this significant increase in subjective desire-satisfaction, Glum finds that he is still just as depressed as ever - that his overall condition has, far from improving, remained roughly the same. Although Feldman doesn't explicitly describe the situation this way, Glum directly parallels the Phenumb character which Carruthers introduced in his 1999 paper: someone who feels neither exultation nor phenomenal frustration when they find that their desires have been fulfilled or thwarted. And what does Feldman say - indeed, think obvious - about the satisfaction/frustration of such a character's desires? That they are worth nothing, or at least very little, when compared to the desires of an ordinary subject. What is perhaps most notable, however, is how the theorists who Feldman's example was targeting (those who uphold preference-satisfaction theories of happiness) have responded to it. They have done this, not by denying that the satisfaction and frustration of Glum's (or, just as well, Phenumb's) desires are relatively unimportant, but by arguing that a person only counts as desiring a thing to the extent that the satisfaction of their desire would actually feel satisfying to them - i.e. by claiming that strength of desire ought to be measured phenomenally, not in a cognitive way. Of course, this path is not open to Carruthers, whose view is that the phenomenal component of desire is ethically irrelevant, and so he is put in the uncomfortable position of having to bite a bullet that nobody else is willing to: holding that we have as much reason to assist Glum or Phenumb fulfil their desires (even though they won't make a felt difference) as we do to assist a regular person fulfil theirs. As before, this seems clearly wrong.

In the above discussion, we have covered pain, negative emotion, and desire frustration. Since this canvasses the major ways in which a moral agent can interfere with another entity, and

since in each case we concluded that the form of interference constituted a greater ethical wrong if the entity being interfered with was conscious rather than lacking any conscious status whatever, this leads us to the conclusion that there is, in fact, a strong connection between conscious status and moral status. More precisely, it leads us to claim we made at the outset: that in order to determine an entity's moral status, we must first be able to determine its conscious status.

SECTION II - EMPIRICALLY INACCESSIBLE CONSCIOUSNESS

The premise of section I was that we can't determine an entity's moral status until we first determine its conscious status. The premise of section II is that the conscious status of WBEs, in particular, can't be determined by the usual combination of empirical study and inference to the best explanation, and so is - in this sense - empirically inaccessible. The specific argument is that there is a fundamental and intractable disagreement over what counts as evidence of consciousness in non-human entities, and that the conscious status of WBEs hinges entirely on the outcome of that disagreement.

Distinguishing Questions: the What, Where, and Why of Consciousness

Whether WBEs are conscious is really just one branch of the broader 'distribution question' of consciousness, which asks where in the universe consciousness happens to arise. A closely related and probably equally ethically important question is the 'phenomenological question' of consciousness, which asks where specific kinds of conscious experience (e.g. valent experiences, or visual experiences) happen to arise in the universe, however this will be largely be put to one side in the following discussion due to time and space constraints. What is important to note, however, is that both of these questions concern the 'where' of consciousness.

A quite separate question is the 'what' of consciousness - i.e. the matter of what phenomenon the word really refers to; of what consciousness *is*, ontologically speaking (this is often termed the 'hard problem' of consciousness, following Chalmers (1996)). To see why this is largely separate from the 'where' of consciousness, notice that the two most popular answers to the hard problem - materialism on the one hand and dualism on the other - are each compatible with nearly any answer to the 'where' question. For example, both functionalist materialists (e.g. Brian Loar, Sydney Shoemaker, Fred Dretske, Stephen Yablo) and naturalistic dualists like David Chalmers hold that consciousness in naturally possible worlds arises exclusively in physical systems with a certain kind of causal structure; the only difference is that the former theorists assert that consciousness ***just is*** the instantiation of that causal structure, whereas Chalmers believes that consciousness is something entirely separate and just happens to naturally supervene upon such structures[24].

Another quite separate question is 'why' consciousness should arise in the physical systems that it does in fact arise in - i.e. the problem of explaining why certain kinds of things (e.g. brains) are conscious and others (e.g. stones) are not. This is the famous "explanatory gap" that Levine

[24] While one's broad answer to the 'what' question has no bearing on the answers one can give to the 'where' question, the same is not true for more specific answers. For example, if you are a materialist functionalist (rather than just a materialist), you would at least have to be committed to the view that all physical systems which have the same causal structure as the human brain (which we know is conscious) are also conscious. This isn't a complete solution to the 'where' question, but it does put limits on what answers we could provide to it. The reverse is clearly also true - if we knew that two systems with identical causal structures had different conscious status, then physicalist functionalism could not be true.

(1983) named. The question of 'why' is closer to 'where' than to 'what', since an answer here (i.e. an illuminating explanation of what it is about brains, rather than stones, that gives rise to consciousness) would likely also give us the conceptual resources to determine which actual entities were conscious, and so fix the answer to the 'where' question. However, almost every answer to the 'where' question still leaves the 'why' of consciousness as quite a mystery. Consider, for example, that even if we somehow came to know, beyond a shadow of a doubt, that only certain biological neural structures were conscious (perhaps only lumps of gray matter like our own brains), this would still leave wide open the question of why those sorts of systems, and not others, were conscious[25].

That the 'where' question - the one we're interested in - can be separated from the 'what' (the hard problem) and the 'why' (the explanatory gap) should be a source of encouragement, since these other questions are among the oldest and most vexed of metaphysical quandaries, no decisive solution to which seems in sight. If an answer to the 'where' really did have to wait on answers to these, many of us might be inclined to abandon the search entirely. In fact, the recognition of this separation was, according to Seth (one of the leading neuroscientific enquirers into the 'where' question) one of the main drivers of the new scientific interest in the subject - as he says "Perhaps the key factor in the transition to scientific legitimacy was the realization that it may not be necessary to explain *why* consciousness exists in order to begin to unravel the physical and biological mechanisms that underlie its various properties" (2010).

Consciousness in Adult Humans: Conscious Correlates

In adult humans, the 'where' question really concerns a handful of marginal cases. We're fairly confident about the conscious status of most people most of the time (we don't wait on the result of science or philosophy to tell us that someone who's going about their day is conscious, nor that someone who's asleep and whose eyes aren't moving is unconscious), but there are some people, on some occasions, who we're not so sure about - e.g. it's not entirely clear whether people are conscious when they sleepwalk, or during memory-loss seizures.

What separates the 'where' of consciousness from any number of other 'where' questions (e.g. where is H2O?) is that consciousness is, seemingly uniquely, a private phenomenon. We cannot peer into the heads of other entities to see what's going on inside, and nor do we possess any such thing as an "expereince-meter"[26] that could do this peering for us. The best we seem able to do in answering the 'where' question is to identify the observables that correlate with our own conscious episodes, and use these to infer the presence of consciousness in others (with the strength of the inference corresponding to the number and reliability of the correlates present). It is for this reason that the science of human consciousness (i.e. the scientific attempt to answer the 'where' question in humans) has often been termed "a science of correlations" (Chalmers

[25] Kathleen Akin puts the matter neatly when she asks: "how is it possible … [that a] gray, granular lump of biological matter, could be the seat of human consciousness?" (1993, p. 124).
[26] An idea attributable to David Chalmers (1996, p. 98).

2014).

Among the thus-far discovered correlates of consciousness, there are two broad kinds: neural and behaviouro-cognitive. The neural correlates of consciousness (henceforth NCCs) are defined as "the minimal neural mechanisms that are jointly sufficient for any one conscious percept, thought or memory, under constant background conditions" (Tononi & Koch 2015). Although no final set of NCCs has yet been determined (many have been proposed[27]), the search thus far has succeeded in narrowing down the possibilities to some form of "widespread, relatively fast low-amplitude interactions in the thalamocortical core of the brain" (Seth et al. 2005, p. 124). This is the kind of activity that doctors look for in EEG scans to distinguish genuinely vegetative patients from merely completely paralyzed ones.

The behavioro-cognitive correlates of consciousness are the set of physical and mental actions which humans only seem able to perform when they're conscious of a certain stimulus. The first and foremost of these - the so-called gold standard of behavioural consciousness studies[28] - is the capacity for "accurate report", i.e. the ability to describe *that* one is conscious of something, as well as what its features are. When a person exhibits this capacity (e.g. by detailing a visual scene in front of them), we generally take that as conclusive evidence that they are conscious of the stimulus in question, and hence conscious more generally. Other key behavioro-cognitive correlates include the ability to durably retain information for immediate use (i.e. to 'hold on' to information), the ability to plan and execute novel strategies, and the ability to integrate temporally separated signals in the service of one's behaviour[29].

It is important to note that, at least as far as the behavioral component of the aforementioned behavioro-cognitive capacities is concerned (i.e. insofar as the cognitive capacities manifest in observable behaviour), the correlation between them and consciousness seems to be an entirely one-way affair. That is to say that consciousness seems always to track these kinds of behaviour, but they do not always seem to track it, as is amply demonstrated by the fact that we can have vivid conscious experiences even while failing to exhibit any behaviour at all (as when dreaming, paralyzed, or simply choosing to be still). The same is not true of neural activity in the thalamocortical regions (out of which we hope, soon, to extract specific NCCs). This activity appears to both accompany and be accompanied by consciousness in every instance we're aware of. In fact, the strength of the neural correlation is such that the absence of such activity, rather than merely constituting an absence of evidence for consciousness, is routinely taken as evidence of an absence of consciousness, at least in adult humans - it is, for example, how we determine whether seemingly comatose patients are vegetative or just non-communicative[30].

---

[27] Chalmers cites at least twenty, referring to the ever-multiply list of proposed NCCs as the "neural correlate zoo" (1997a, p. 1)
[28] It is explicitly referred to as such in (Tononi and Koch 2015).
[29] The first two of these correlates are raised in (Dehaene & Naccache 2001, pp. 9-10); the third is discussed in (Seth 2016, p. 3).
[30] As Martha Farah notes (2008, p. 13), "functional neuroimaging [has provided] a new window on the mental status of severely brain damaged patients".

The apparent non-contingency of the neural-consciousness relationship, as compared to the clear contingency of the behaviour-consciousness relationship (though perhaps not the cognitive-consciousness relationship), has led some to urge that we try to move away from behavioural evidence altogether, e.g. Farah (2008). This, however, does not seem warranted; if our purpose is simply to identify instances of consciousness, rather than to arrive at any deep metaphysical conclusions about what consciousness really *is*, a one-way correlation should be perfectly sufficient. As proof of this, consider that it would never seem sensible, having gotten an accurate experiential report from a person, to say that we should wait on their brain scan results before coming to a final verdict as to their conscious status.

Consciousness in Non-Human Animals: Analogues and Synthesis

When, if ever, are animals aside from adult humans conscious? Many animals certainly seem conscious - it's hard to doubt that a puppy scrabbling at the door when it hears you coming down the drive isn't really feeling happy, or to imagine that an octopus might have perceiving eyes yet lack a visual field, and even insects have often struck people as possessing an inner life something like our own[31] - but how much of this is quixotic anthropomorphism and how much is fact? To answer this question, most agree that the surest approach is to search for analogues of the correlates of consciousness in humans. Unfortunately, the analogues we find in other species are often imperfect and difficult to identify, which adds a much higher degree of uncertainty to our corresponding ascriptions of consciousness (this being one of the chief reasons why the 'where' question of animal consciousness is still, to a quite large degree, unsettled[32]).

Consider, for example, how we might identify animal analogues of our 'gold standard' behavioural correlate of consciousness: accurate report. Most animals obviously lack the capacity for verbal communication, but this doesn't rule out the possibility that they might make detailed behavioural reports of their experiences in the same way that, for example, human infants can still communicate specific information about their experiences without the use of words (e.g. if a toddler bursts into tears after we take away its toy, and continues in earnest when we hand it back a slightly different one, we can be fairly confident it's noticed the difference). The trouble is we know that, even in humans, much behaviour which seems to evince an awareness of a particular object or piece of information actually occurs without any awareness of it at all (e.g. you can prime people's responses with stimuli they can't consciously see (Naccache et al. 2002), and can get certifiably brain-dead human patients to track visual stimuli with their eyes[33], not to mention the broad suite of behaviours that are performed - likely

[31] Charles Dickens, for example, writes in Great Expectations that "the black beetles ... groped about the hearth in a ponderous elderly way, as if they were short-sighted and hard of hearing, and not on terms with one another".

[32] Giulio Tononi, another prominent neuroscientist working in the field, going so far as to say that "no consciousness expert, if there is such a job, can be confident about the correct answer to such questions" (2008, p. 217).

[33] As Farah (2008, p. 11) notes: "vegetative patients may move their trunks and limbs spontaneously, and have been observed to smile, shed tears, and vocalize with grunts. They may even orient their eyes and heads toward peripheral visual motion or sounds".

unconsciously - by somnambulists).

The challenge is to be able to distinguish animal behaviour that indicates genuine conscious awareness of a stimulus from that which merely indicates "the ability to [potentially unconsciously] distinguish between, and generalize across, classes of stimuli" (Seth et al. 2005, p. 120). We can figure this out experimentally in the human case, but when it comes to other species it's very difficult to know where the line lies. To the extent that we can know this, our knowledge tends to be based on similarities between the animal's behavioural and neural profile and our own. For example, it is now widely accepted that macaque monkeys can provide accurate reports about their visual experience by pressing specific buttons on a touch-screen, but this is only accepted because "the macaque visual cortex has striking similarities to that of the human" (Ibid., p. 120), and because their button-pressing behaviours respond to brain lesioning in essentially the same way that human accurate report does (e.g. removal of the striate cortex causes button-pressing behaviour reminiscent of blindsight[34]). As we move beyond primates to consider more evolutionarily distant species, these similarities drop off, with the result that accurate report becomes much harder to identify, and correspondingly less useful as an indicator of consciousness (in stark contrast to the human case, where it was something of a silver bullet).

The same interpretative challenges come into play with our neural correlates. We know - or hopefully will know soon - what neural structures and dynamics underlie consciousness in adult humans (these being the NCCs), but finding perfect analogues of those features in other creatures is often impossible. Instead, we search for neural mechanisms that are similar to our own in ways that seem important. In particular, we search for analogies of structure (e.g. the composition of different brain parts and the way they are connected), and analogies of dynamics (e.g. the way different parts operate and interact with each other over time), with each of these categories further divided according to the scale at which the analogy occurs, i.e. whether the similarity appears at the neuronal, circuit, or network level. If we can identify a neural mechanism in some other animal that is analogous to one of our own NCCs both structurally and dynamically at a number of different levels, and if this mechanism can further be shown to have originated in a common evolutionary ancestor (in which case these neural similarities will be homologies, rather than mere analogies), then we are generally fairly confident in ascribing consciousness to the creature. Of course, as homology gives way to analogy, and as the number and strength of these analogies drops off, uncertainty seeps in once more.

Because any answer to the 'where' question with respect to animals must rely centrally on locating analogues of the human correlates of consciousness, and because most analogues are imperfect and can only support a fairly weak inference of consciousness on their own, researchers usually try to identify as many of these analogies as possible and, in particular, to identify analogies of both behavioural and neural kinds. It is precisely this sort of evidential synthesis that grounds the nearly unanimous ascription of consciousness to mammals (whose

[34] This being the famous result which Cowey and Stoerig (1995) showed.

brains and behaviours are both incredibly similar to our own), and which is now being used by researchers to try and argue for avian consciousness. Indeed, some have gone so far as to assert that only a synthetic approach can suffice, and that "a compelling case for avian consciousness cannot be made solely on the strength of relevant neuroanatomical and neurophysiological resemblances … [nor on] avian behaviors that imply sophisticated cognitive capabilities" (Edelman & Seth 2009, p. 481).

When Correlates Collide

The lack of perfect analogues of human correlates of consciousness is certainly one challenge in investigating animal consciousness, but there is another that could potentially be even more troublesome - a problem which in principle could thwart, rather than just muddy, the inquiry into the 'where' question. This is the prospect that the behavioural evidence and the neural evidence might actually diverge in some cases - i.e. that certain animals might at the same time exhibit behaviours which, in us, would strongly indicate conscious experience while sporting a neural system that is disanalogous to our own at essentially every level.

This is not a purely academic worry, and indeed seems to be at the heart of the heated scientific debate currently occurring around the conscious status of fish and, more specifically, over whether fish consciously feel pain. All parties to this debate agree that fish display aversive behaviour towards noxious stimuli (so-called 'nocifensive' behaviour), but this is not on its own sufficient evidence that they consciously feel pain, since we know that a large number of nocifensive behaviours are exhibited in humans and other mammals even in the absence of conscious pain - for example, humans under general anesthetic will still jerk away from the touch of a scalpel unless they are given local muscle relaxants (this movement being controlled entirely by the spinal cord and brainstem), and decerebrate rats (who have their entire brain above the midbrain removed, including all the apparatus necessary for producing felt pain in mammals) react nearly indistinguishably from regular rats when receiving an injection, often wriggling, vocalizing, and trying to bite the syringe (Rose 2014, p. 101). The reason that fish present such an interesting case is that, on top of these sorts of simple, seemingly reflexive nocifensive behaviours (which are at best ambiguous indicators of consciousness), they also seem to display a whole suite of more complex behaviours - behaviours whose analogues, if exhibited by humans, would provide strong grounds for ascribing consciousness. For example, when injected in the lip with bee venom, rainbow trout won't feed for three hours, will rub their lips on the tank's gravel, and often display a strange rocking motion that somewhat resembles the rocking which primates exhibit when in pain, and which is thought to be a comfort response (Sneddon 2003). Separate studies show that fish can perform risk reward trade-offs (Milsopp and Lamming 2008), and that they will remember the circumstances that surrounded a traumatic incident and use this information to adapt their behaviour - paradise fish, for example, will avoid locations where a predator attacked them for up to three months (Csanyi et al. 1989), while carp will avoid human bait for up to three years after being hooked just once (Beukema 1970). What is important about these example behaviours is that they seem to involve

novelty and informational integration - cognitive capacities which, recall, are key correlates of consciousness in humans.

From a behaviouro-cognitive standpoint, then, the case for conscious fish pain seems strong. Unfortunately, when we open up the skulls of fish to search for corroborating neural evidence, we find essentially none. In particular, we find that fish appear to lack any "readily identifiable homologs of neural mechanisms associated with mammalian consciousness" (Seth 2016, p. 3) and, more specifically, that they lack any plausible analog of our human cortex - a brain region that is essential for human consciousness in general, and for conscious pain specifically. It's difficult to overstate the relevance of this particular disanalogy to the dispute over fish pain: in mammals, the cortex is not just one important link in a chain of brain-parts responsible for generating felt pain; it is, as best as we can tell, the sole seat of the experience of pain. The cortex is the area of our brain that lights up under EEG scans when we report pain (but not when we're under the effects of an anesthetic); that, when lesioned, removes or interferes with this feeling while keeping basic nocifensive reflexes intact; and that, when remotely stimulated, can cause us to report pain even in the absence of any nerve activation. So close is the connection between cortical activity and phenomenal pain that algorithms can actually be trained to accurately predict the pain reports of human patients exclusively from cortical brain scans. It is no small matter, then, that fish brains seem to lack any component that is even grossly functionally analogous.

The battle lines in the debate over fish consciousness are essentially drawn around these two opposing strands of evidence. On the pro-pain side - which has the preponderance of voices - it is argued that while it might have increased our confidence that fish were conscious if we had found an abundance of corroborating neural evidence, the absence of such evidence should not count as evidence of an absence of consciousness, especially since we're not yet sure exactly what it is about mammalian brains that allows them to generate consciousness in the first place (this being the notoriously tricky 'why' question of consciousness). Proponents of this view point out that convergent evolution often produces different structures capable of supporting the same high-level functions - consider, for example, the many separate instances of wings having evolved (e.g. insects, feathered birds, bats), or the more than a half-dozen separate evolutions of focusing eyes - and that fish brains might just represent an alternative neural realization of a conscious system. They further argue that, if evolution had produced such an alternative system, the way to identify it would be to search for the telltale "cognitive and behavioural repertoires" (Ibid., p. 3) that consciousness seems to enable in us - i.e. the kinds of behavioro-cognitive capacities that we do, in fact, find in fish.

On the anti-pain side, scientists like James Rose (2002; 2014) and Brian Key (2015; 2016a; 2016b) point out that the behaviours which, in humans, happen to require consciousness are not divided from those which can be performed unconsciously by any clear or fundamental principle - it is not the case, for example, that the latter are all simple and reflexive while the former are all complex and coordinated (consider, for example, the highly complex 'righting reflex' or the often sophisticated behaviours of somnambulists). Indeed, the boundary line between conscious

and unconscious human behaviour has, for the most part, had to be determined by meticulous experimentation rather than any internal coherence - take, for example, the discovery that trace conditioning[35] requires consciousness, while regular Pavlovian conditioning does not. Since this is the case, they argue, it seems perfectly plausible that a species which had never evolved any capacity for consciousness might develop sophisticated unconscious methods of performing many of the tasks that mammals like ourselves rely on consciousness to perform. For this reason, they argue, it won't suffice to simply search through the animal kingdom for instances of our own conscious behavioro-cognitive correlates; one must also locate "a plausible neural mechanism" (Rose 2014, p. 120) underlying these capacities that might sustain consciousness. This is not to rule out the possibility of consciousness being multiply realized by convergent evolutionary processes; it is simply to insist, as seems only sensible, that every conscious system must have at least some fundamental properties in common (in the same way that all focusing eyes must at least have a lens, an aperture, and some photoreceptor cells, even if these can be constructed from very different materials), and to note that, in humans at least, the only thing consciousness seems to be fundamentally and non-contingently tied to is a particular sort of neural activity.

The disagreement between these two sides is perhaps best framed using the language of multiple realizability. On the one side, those open to fish pain conceive of consciousness as a high level behavioro-cognitive phenomenon capable of being realized by a diverse range of neurally distinct systems (from humans to fish, and maybe even to octopuses and goopy green aliens). On the opposing side, consciousness is conceived of instead as a lower level neural phenomenon with a much narrower scope for realization (being instantiated perhaps only by mammals and birds). The former accuse the latter of unjustifiable neural anthropocentrism; the latter accuse the former of treating the brain as a "mysterious black box" (Key 2016b, p. 3) whose only relevant properties are its gross behavioural outputs. The million dollar question, of course, is which group is right.

A Problem of Bridging Principles

The problem of fish consciousness might, of course, end up being solved without the deeper disagreement needing to be. We might, for example, discover that there actually are some deep neural analogies or homologies between fish brains and mammalian ones that previous research had failed to uncover (a discovery that would be very in keeping with neuroscientific trends over the past few decades, and which at least some parties to the debate are optimistic about), so that the neural evidence and the behavioural evidence cease to be in tension. What is important to note is that even if this were to happen, the fundamental disagreement would still stand in need of resolution, not least because it is exactly the disagreement that the conscious status of WBEs (the specific kind of AI this thesis is centrally concerned with) depends on.

[35] Trace conditioning requires an entity to learn to associate two temporally separated stimuli, whereas classical Pavlovian conditioning has the conditioned and unconditioned stimulus occur simultaneously.

If consciousness can be inferred solely from the exhibition of certain behaviouro-cognitive capacities, then WBEs will have as strong a case for consciousness as any human being would (being behaviourally identical). However, if consciousness is deemed to be essentially a product of neurology - the brain being, after all, the seat and engine of consciousness - then the case is much less clear. On the one hand, WBEs have identical neurology at the circuit and networks levels of abstraction, but on the other hand both their composition and their function at the sub-neuronal scale is completely unlike that of a regular brain - WBE neurons are virtual entities with no real internal detail[36], whereas human neurons are concrete entities containing rich systems of biochemical processes.

What exactly is this evidential disagreement about? *Prima facie* it might seem to be a metaphysical dispute about the fundamental nature of consciousness (i.e. whether it is fundamentally a cognitive or a neural phenomenon), but this is not so. The two sides are not divided over the logical or metaphysical essence of consciousness, but merely over its natural essence - how it manifests in the actual world, rather than how it must manifest *simpliciter*. Neither side in this debate is committed to a claim that consciousness is identical to a certain sort of physical system or property (such "psycho-physical identity statements", as Levine called them (Levine 1983, p. 354), are the stuff of the hard problem); all they are committed to are claims of natural supervenience, i.e. claims that the facts about the instantiation of certain physical systems or properties (whether they be neural or behavioural) happen to fix the facts about the instantiation of consciousness in the real world (saying nothing about other logically possible worlds).

These sorts of supervenience claims are what David Chalmers (1996, p. 218) refers to as "bridging principles". They are principles that link conscious experience to certain physical systems in the natural world, so that we can identify the former by identifying the latter (the latter being something that, unlike consciousness, we can in theory directly observe). Such principles, says Chalmers, are not arrived at by performing experiments on creatures with unknown conscious status (since, without a bridging principle already in hand, we'd have no way to interpret the results of such experiments). Rather, they are arrived at by studying regularities between the conscious data we're most confident of (i.e. our own first-person experiences, and the reported experiences of other humans) and the physical processes underlying them, then asking ourselves what lawlike principle would most simply and plausibly explain these regularities. For example, we might note that whenever we - or a fellow human - is conscious of some information, this information appears to be directly available for verbal report as well as for guiding our voluntary behaviour (Chalmers calls this behavioro-cognitive property "global availability" (Ibid., p. 223), and that the reverse also appears to be true: wherever some information in our environment is directly available for us to report on and to be used in directing our behaviour, we are generally conscious of it. This is in fact the regularity

[36] Bostrom (2014, p. 40) mentions that, at least in principle, a WBE could accurately simulate a brain down to the atomic level using the quantum-mechanical Schrodinger equation, but that the only WBEs likely to be realistically produced will have individual neurons as their basic building blocks, along - perhaps - with certain subsystems like dendric trees and synapses.

which Chalmers focuses on, and which leads him to propose - as his candidate bridging principle - that consciousness is instantiated in the natural world wherever we find "the use of information in deliberate and controlled behaviour" (1997b) spanning "many motor modalities" (1997a, p. 4).

Chalmers anticipates that we might encounter a problem if we find "two equally simple theories [i.e. bridging principles] both of which fit the [conscious] data perfectly" (Ibid., p. 201), since in this case we would seem to have no criteria to choose between them. However he does not see neural bridging principles as raising this possibility, because in his view such principles essentially propose that consciousness will arise wherever we find some behaviouro-cognitive property (e.g. global availability, or any of the behavioural correlates mentioned earlier) ***plus*** some additional neural "X-factor" underlying the behavioro-cognitive property (e.g. a flesh and blood thalamo-cortical system, or some neural analogue). According to Chalmers, adding this extra neural ingredient doesn't make the principle fit our conscious data any better than it otherwise would (behavioural properties already achieve a perfect fit, he thinks, since that's how we picked out our conscious data in the first place); all it really does is needlessly complicate the story and so, like any explanatorily impotent appendage to a scientific theory (e.g. the luminiferous aether which physicists once postulated as a medium for light to travel through), it should be abandoned.

This does not seem to me to be a convincing argument, at least not against every type of neural bridging principle. To see why, consider the family of neural bridging principles which claim that a system can only be conscious if it exhibits seemingly conscious behaviour (e.g. the aforementioned behavioro-cognitive correlates of consciousness) ***and*** where that behaviour is generated by a neural system which is compositionally and functionally analogous to our own at every scale of abstraction, all the way down to the molecular. Henceforth, this sort of principle will be referred to as a 'substrate neural principle' or SNP. If some kind of SNP is correct, then a system such as a WBE would not count as conscious since its functioning at the sub-neuronal scale (i.e. the internal structures and dynamics of its neurons) differs dramatically from our own. In fact, even a WBE that modelled a human brain accurately all the way down to the level of individual atoms (something computationally impracticable, but not in principle impossible) would still not qualify for consciousness according to SNPs, since at this finest of scales there would remain a deep compositional disanalogy between the WBE and a human brain - the one being made of virtual atoms, and the other being made of actual ones. Let us examine how this kind of principle weathers Chalmers' criticism.

One part of Chalmers' objection - that a principle like this would be more complex than any purely behavioro-cognitive bridging principle - certainly hits the mark. In fact, SNPs would likely be the most complex form of neural bridging principle and so especially open to this criticism[37]. The important question, however, is whether this added complexity is truly needless,

[37] A substrate neural principle would not, of course, need to specify exactly what a physical system be composed from, or how it need operate, to be conscious - this would be overkill - but in demanding some

as Chalmers claims, or whether it affords certain explanatory advantages that would be lost if we stripped these neural principles back to just their behavioro-cognitive conditions. I think a close inspection reveals the latter to be the case: the complexity of SNPs actually affords them at least two important theoretical advantages - advantages, what's more, that at least plausibly outweigh the disadvantage of their complexity. These are:

I. ***Better fit with the conscious data.*** On Chalmers' view, a bridging principle's purpose is to explain the correlations we observe between our conscious data and physical processes. The conscious data itself, he thinks, consists of our first-hand conscious experiences as well as the accurately reported experiences of other humans. The justification he gives for including the reported experiences of other humans as paradigm conscious data is that, if such reports turned out to be unreliable indicators of consciousness (i.e. if philosophical zombies were a possibility in the actual world) then "all bets would be off … and a theory of consciousness would be beyond us" (Ibid., p. 200) - this is, in effect, a sort of methodological grounds for treating such data as reliable.

   These two sets of conscious data (our first-personal data and the reports of other humans) do not, however, seem to exhaust all of the data relevant to deciding between bridging principles - in fact, they seem just to be the most certain of our data points. Consider, for example, the many seemingly conscious behaviours that various mammals exhibit. Unlike with human verbal reports of consciousness, there are no methodological grounds for treating these mammalian behaviours as reliable indicators of consciousness, since the discovery that they were not would be perfectly compatible with further consciousness research going on - it would not mean that "all bets were off". In fact, the discovery that non-human mammals lacked consciousness entirely would be a great boon for consciousness research, since it would narrow down the physical processes essential for consciousness to those which humans possessed but which other mammals lacked (e.g. language use or the systems underlying self-consciousness). But this is the important point: despite the fact that there are no reasons, methodological or otherwise, to treat mammalian consciousness as certain, most of us nonetheless have a very strong intuition that some mammals are conscious at least some of the time (e.g. consider again the case of the puppy scrabbling excitedly at the door). This seems like an intuition that a proposed bridging principle would do well to endorse. Indeed, if two bridging principles differed only in their verdict on mammalian consciousness (i.e. if both were able to account for the human consciousness data equally well), this alone would seem to provide a strong reason to prefer the principle that conferred consciousness on mammals. This is not to say that any bridging principle that denied mammalian consciousness should be discounted (we aren't imposing a strict plausibility constraint), but rather to say that there would have to be strong countervailing reasons

level of analogy to the human brain at every scale of abstraction it would still be introducing a great deal of complexity

to support such a principle (e.g. perhaps if it were the only principle that provided a reasonable explanation of the human consciousness data). To sum up, then: we have more conscious data than we have human conscious data, even if the human data is our most indubitable and what we might call our "primary data"[38].

The relevance of this fact to SNPs is the following: most people do not just have strong intuitions that some mammals are conscious some of the time; they also have very strong intuitions that certain kinds of physical systems (e.g. stones) lack consciousness entirely. Of course, neither behavioro-cognitive nor neural bridging principles ascribe consciousness to stones, but there are some hypothetical physical systems which would qualify as conscious on both behavioro-cognitive and higher-level neural bridging principles (i.e. all neural bridging principles bar SNPs) even though it seems intuitively obvious that they are not. A key example of this sort of hypothetical system comes from Ned Block's 'Chinese nation' thought experiment (1978, p. 239). In this, he imagines what would happen if the government of China decided to organize its citizens (whose number is comparable to the number of neurons in the human brain) in such a way as to enact the functional structure of an actual human brain. Each citizen would represent a neuron or some other microscopic brain component, and would be given a walkie-talkie to communicate with certain other components. They would also be given a look-up table which would tell them, for any conceivable set of input communications, what output communications to make - instructions of the form "If walkie-talkie 391 calls, then call walkie-talkie 615" (Prinz 2012, p. 282). Where a component of a the original biological brain would send signals to some other body part (e.g. to produce a motor response), the equivalent citizen-component of the nation-brain instead sends signals to transmitters in a humanoid robot, which performs the equivalent action. Likewise, where the components of the original brain would ordinarily receive input from nerves in a flesh-and-blood body, the citizen-components of the nation-brain receive inputs from analogous sensors in the aforementioned robot. The overall idea is that this robot, which both receives information from and feeds information to the nation-brain, would - if everything were set up correctly (which is at least possible in principle) - behave in just the same way as the actual human body whose brain organization the Chinese nation was enacting.

Block thinks it obvious, and indeed it does seem intuitively obvious, that neither this collection of Chinese citizens nor the hollow robot which it controls would have any consciousness of its own (though obviously the citizens themselves would all be conscious). Yet it is clear that since the robot in question is behaviourally indistinguishable from a conscious human, and since the network of Chinese citizens has the same neural structure a conscious human brain (at least at a super-neuronal level of abstraction), both a behavioro-cognitive and a higher-order neural bridging principle would have to ascribe consciousness to such a system. What is equally clear is that a SNP

[38] This term is borrowed from Dehaene and Naccache (2001, p. 3).

would deliver the intuitively correct result here, since the Chinese nation's neural structure is - at a sub-neuronal scale of abstraction - deeply functionally and compositionally disanalogous to that of a normal brain (the internal details of a Chinese citizen being radically different from that of a regular neuron, even in the case where their input-output behaviour is made to correspond).

Some, e.g. Jesse Prinz, find Block's hypothetical and similar thought experiments wholly unconvincing because "there is no reason to think that these intuitions [about what systems can and can't be conscious] track reality" (Ibid., p. 282). As Chalmers himself notes: "it is equally intuitively implausible that a brain should give rise to experience! Who ever would have thought that this hunk of grey matter would be the sort of thing that could produce honest-to-goodness experiences" (Chalmers 1996, p. 235). These responses would be decisive if the Chinese nation example had been presented as knock-down counterexample to behavioro-cognitive and higher-order neural bridging principles, but - while this may have been how Block originally presented it - this is not how it is presented here. We aren't arguing that our intuitions about systems like the Chinese nation should count as *primary* conscious data (in the same way that our first-person conscious experiences do); we are simply arguing that they should count for something, in the same way that our intuition that door-scrabbling puppies are conscious should count for something. This proposition seems perfectly reasonable - after all, I am about as sure that a nation of walkie-talkie equipped Chinese citizens would not give rise to an independent consciousness as I am that many mammals are conscious - and it is also completely immune to the objections raised by Prinz and Chalmers. Insofar as all this is correct, then, it seems to count in favour of SNPs that they deliver the intuitive result in hypothetical cases like these. What's more, since such principles achieve the same fit with the rest of our conscious data (e.g. our human and mammalian conscious data), it would seem that, contra Chalmers, their complexity does afford them some theoretical advantage, namely a better fit with our data (the coin of the realm with scientific theories).

II. ***Increased predictive precision.*** While the question of what counts as an analogue of some particular neural mechanism is often just as fraught with vagueness and interpretive difficulty as parallel questions concerning behavioural analogues (the business of identifying analogies being inherently imprecise, as discussed earlier), this paradoxically becomes less and less a problem the more analogues we are simultaneously searching for. Consider, for example, the question of determining whether songbird behaviour indicates a capacity for consciousness (leaving neurology aside). On the one hand, there is genuine uncertainty surrounding the matter of whether songbird vocalizations are analogous to our human verbal reports[39] (this being a key behavioural correlate of consciousness, as mentioned already), however this uncertainty does not muddy the overall question because we can identify clear analogues of a number

[39] This question is the subject of the discussion in (Edelman & Seth 2009, p. 479)

of the other behavioural correlates of consciousness in songbirds and we take them to suffice. It is a more general truth that if the clear presence or absence of a single property can be decisive in settling an entity's conscious status, then the ambiguity around each individual property matters less and less as the overall number of properties increases. After all, the only way the uncertain nature of individual properties could bleed through into uncertainty about a creature's overall conscious status in such a case would be if that creature happened to fall into the interpretive 'grey areas' of each and every property (a possibility whose likelihood shrinks as the number of properties multiplies).

Take the example of David Chalmers' proposed bridging principle, which identifies just one behavioro-cognitive property as relevant to conscious status ("global availability" - i.e. the direct availability of information to all a creature's motor modalities). Since even Chalmers admits that this property has vague boundaries (it being "only clearly defined for cases approximating human complexity" (1996, p. 230)) the acceptance of his bridging principle would render the conscious status of vast numbers of small and microscopic organisms effectively indeterminate. The same problem is raised, to a lesser extent, by any behavioro-cognitive bridging principle which specifies only a small list of capacities as relevant for identifying consciousness - there is a non-negligible risk that some entity might only exhibit one or two of these, and then only ambiguously, in which case its overall conscious status will be profoundly uncertain. A SNP drastically reduces this risk, since it stipulates - as a further condition for conscious status - that an entity also manifest neural analogies at a whole range of scales of abstraction (from the network-level down to the molecular). In effect, such a principle adds a whole suite of failure conditions to the small pool of success conditions that behavioro-cognitive bridging principles already stipulate. In this way, it offers far more predictive precision - another key theoretical virtue.

Where does this leave us? Well, if we agree with Chalmers about how bridging principles are justified (that they are based on paradigm-case consciousness data plus inference to the best explanation), but disagree with him insofar as we hold that substrate neural principles are not obviously less plausible than behavioural ones (faring better by some standards of theoretical plausibility, but worse by others), then we seem to be in just the predicament he worried might arise - namely of having two conflicting rules for inferring consciousness from physical evidence, and no criteria for deciding which is correct. Since the conscious status of WBEs depends critically on which of these principles is right (WBE consciousness being guaranteed by any behavioro-cognitive bridging principle, but being ruled out by an important subset of neural bridging principles), it follows that the conscious status of WBEs is empirically inaccessible.

SECTION III - ARGUMENTATIVELY INACCESSIBLE CONSCIOUSNESS

The fact that evidence and inference to the best explanation alone are not sufficient to settle the dispute between behavioural and neural bridging principles does not mean the dispute has no solution. In particular, it seems entirely possible that a process of imaginative argument might succeed in adjudicating the matter where more empirical methods failed. This is precisely the approach David Chalmers takes with his 'fading qualia' argument (1996, 236; 2010, 37). The argument is not designed to completely settle the neural-behavioural dispute, but it would - if successful - eliminate the particular subset of neural bridging principles which deny consciousness to WBEs (those termed 'substrate neural principles' in section II, and henceforth referred to as SNPs), thereby settling the dispute insofar as it pertains to this thesis. Recall that SNPs were those neural bridging principles that place sufficient weight on microscopic neural composition and functioning that an entity such as a WBE (which has identical functioning to a human brain at a neuronal grain of abstraction, but is dramatically disanalogous in both composition and function when considered at finer grains than this) would fail to qualify as conscious. As an example of this sort of principle, Chalmers mentions the seemingly intuitive proposal that only neural systems based on "cell-based biology" could be conscious - a proposal he calls the "biological" theory of consciousness (2010, p. 36).

Chalmers' argument against such principles takes the form of a *reductio ad absurdum*. He begins by assuming that differences at the substrate level could make the critical difference between an entity's conscious status. More precisely, he assumes "there could be a system with the same functional organization as a conscious system … but which lacks conscious experience entirely", where 'functional organization'[40] is understood as referring to a system's abstract causal structure considered at a neuronal grain of coarseness (i.e. ignoring components more microscopic than neurons and synapses, except insofar as they influence the input-output behaviour of these larger components). If two such "functionally isomorphic" systems existed, their behaviour at every level above the neuronal would by definition be identical, and yet one of them would be - in effect - a zombie system. For the sake of argument, Chalmers asks us to imagine that the two systems do exist, and that they are a conscious human brain and a replica of the same brain - which he calls Robot - made from silicon microchips rather than biological neurons (of course, he could just as easily have supposed that the isomorph was composed of virtual neurons, and so the argument generalizes to WBEs).

Given these two systems, says Chalmers, we can easily imagine a series of intermediate cases between them. The first such case would be created by taking the conscious human brain and replacing one of its neurons with the functionally isomorphic component from Robot. This shouldn't present any in-principle problem because, after all, the two systems have identical

[40] The exact way Chalmers defines this is the "abstract pattern of causal interaction between [the system's] components … [and] between these components and external inputs and outputs" (1996, p. 231), where the system is represented at what he calls a "fine enough grain" (Ibid., p. 238). He later goes on to explain that this is any grain "fine enough to determine the behavioural capacities associated with the brain" (Ibid., p. 232), and uses - in his examples - the neuronal scale.

causal structures at a neuronal level of abstraction, so that every neuron or synapse in the human system has a direct counterpart in Robot that fits into a counterpart network in exactly the same way. In the case of a WBE, whose neurons are virtual rather than physical, this replacement would occur by removing the original biological neuron and attaching transmitters to all its old input-output channels - these transmitters would then send input signals to some external computer, where the neuron would be simulated virtually and the resultant outputs transmitted back. Having replaced one neuron, we could then construct a second intermediate case by replacing two neurons and the synapse joining them, and we could continue in this fashion creating a complete chain of intermediate cases between an entirely conscious biological brain and an entirely nonconscious virtual one (i.e. Robot). Having constructed this chain, Chalmers wonders whether and to what extent each link in the chain would be conscious. As he notes, there are essentially only two options: either consciousness fades gradually as the system becomes less and less biological, or else it disappears suddenly after a single replacement. A third option, of course, is that a conscious capacity remains completely unaffected through the whole process, but this would contradict our starting assumption that Robot is unconscious and so we leave it aside. Of the first two options, only the gradual fading out of consciousness seems like a realistic possibility; the alternative - a sudden disappearance - would require us to believe that a system's entire conscious capacity could hinge on the replacement of a component such as a neuron - a possibility that Chalmers rightly describes as "extremely implausible" (1996, p. 238).

With the above in hand, Chalmers then considers what a system halfway through this transformative process might be like. On the one hand, it would have to have a highly degraded phenomenological experience. It's not clear exactly how this would manifest - Chalmers suggests that experiences which, in a human brain like our own, would be vivid and rich (e.g. a crisp summer scene) might, to this hybrid system, be experienced instead as faint and darkened, perhaps with certain subtle distinctions missing - but it seems clear that some sort of fading would have to occur. On the other hand, the macroscopic functioning of this system, including all its cognitive and behavioural operations, would be identical to that of the original biological system it was crafted from, since the only difference between the two are neuronal-level replacement components and all of these were functionally isomorphic to the original components (i.e. they produce exactly the same outputs given the same inputs). What this means in particular is that the hybrid system would have the very same beliefs, and make the very same reports, about its experiences as would the original biological system, even though in reality the hybrid system's experiences would be vastly diminished. Such a system would be, as Chalmers puts it, "systematically out of touch with its experiences" (Ibid., p. 240).

This disconnect between the hybrid system's judgments about its experiences and the reality of those experiences is the result which Chalmers takes to be absurd, and which powers his *reductio* argument. He does not claim it is a strict logical impossibility - he notes that "there is no contradiction in the description of a system that is so wrong about its experiences" (Ibid., p. 240) - but he believes it is implausible enough to justify abandoning the original assumption that substrate-level neural differences could make an important difference to an entity's

conscious capacity. To bolster this intuition and highlight just how absurd a result this really is, Chalmers provides two further arguments. Firstly, he makes the empirical point that "in every case with which we are familiar, conscious beings are generally capable of forming accurate judgments about their experience, in the absence of distraction and irrationality [he later also adds "functional pathology"]" (Ibid., p. 239). Secondly he notes that, if we suppose these sorts of microscopic changes at the substrate level really could make a significant difference to our conscious experience - a difference, what's more, that we would be unable to notice - then we are led to the worrying thought that such changes "might be actual, and happening to us all the time" (Ibid., p. 254). After all, low-level physiological changes of this sort (e.g. cell replacement) occur nearly continuously in the ordinary human brain. Chalmers thinks that the reason such possibilities don't seriously worry us is that almost everyone takes for granted the principle that "when one's experience changes significantly, one can notice the change" (Ibid., p. 254), and that a consistent application of that principle must rule out the kind of fading qualia scenario described here.

It is important to note, here, exactly why it is that Chalmers rejects the argument from logical impossibility, because on its face this seems to be one of the more compelling reasons for judging the fading qualia scenario absurd. In particular, it seems like one might reasonably secure the absurdity result on the grounds that the very concept of an unnoticed (indeed unnoticeable) phenomenal experience is incoherent[41] - i.e. that "it is a constitutive property of qualia that we can notice differences in them" (Ibid., p. 258). This differs from Chalmers' own claim (i.e. from the principle mentioned above that he believes most of us take for granted) in that it concerns logical impossibility rather than just natural impossibility. Chalmers stops short of this stronger claim for two reasons: in the first place, he just doesn't have the conceptual intuition that underpins it; but in the second place, and more relevant for our purposes, it is entirely at odds with dualism (Chalmers' preferred answer to the 'what' question of consciousness). According to dualism, phenomenal consciousness is a fundamentally non-physical kind of thing, and facts about the instantiation of phenomenal consciousness are logically independent from ordinary physical facts. To use Chalmers' own explanatory metaphor, dualism is the view that God had "more work to do" (Ibid., p. 110), after fixing all the physical facts (e.g. about the positions of subatomic particles), before it became true that certain physical systems were conscious.

The reason dualism is incompatible with the aforementioned claim (that it is constitutive of conscious experiences that they be noticed, or noticeable) is that 'noticing' and 'noticeability' are psychological phenomena which can be entirely reduced to physical systems and processes such as working memory - a fact that Chalmers openly admits[42]. If phenomenal consciousness really

[41] This seems to be the view that Michael Cohen and Daniel Dennett are gesturing at when they asks, rhetorically, "what does it mean to have a conscious experience that you yourself do not realize you are having?" (2011, p. 362)

[42] Specifically, he says "awareness is the psychological correlate of consciousness, roughly explicable as a state wherein some information is directly accessible" (Chalmers 1996, p. 203) and suggests that the two might be connected by a "psychophysical law" (Ibid., p. 201).

were partly constituted by physical things like these, then it could not be the case that physical facts and conscious facts were logically independent of one another. Instead, it would seem that God, in fixing the physical facts, would effectively have restricted the ways in which he might fix the phenomenal facts (since he would not be able to imbue any system with consciousness that he had not already imbued with certain functioning psychological capacities). Likewise, it could not be the case that across all logically possible worlds the term consciousness just picked out the 'stuff' in any given world which had a "phenomenal feel" (this being one of dualism's central claims[43]); rather, the term would only pick out the 'stuff' in any given world which had both a phenomenal feel ***and*** the physical property of being noticed or noticeable. The reason this is important for us is that it means the conceptual incoherence argument is unavailable to anyone who would wish to remain agnostic on the dualism-physicalism debate. Since this includes us (this thesis aiming to avoid such stormy waters), we will therefore leave this line of argument aside in the discussion that follows.

What remain, then, are the two arguments Chalmers did endorse, and the question of whether these really demonstrate the absurdity of the fading qualia scenario with sufficient force to act as a reductio. The first of these arguments, recall, was that the scenario describes a divergence of experience and experiential judgment that is implausible because it is empirically unprecedented (at least among attentive, normally functioning humans). The second was that, if this sort of divergence really could be caused by physical changes at the level of an entity's neural substrate, there would be good reason to think this was occurring to ordinary humans like ourselves essentially constantly - a result which, again, Chalmers takes to be implausible.

Both of these arguments seem to be open to decisive objections. Against the first, it suffices to note that if there ever had been a disassociation between a person's conscious experiences and their judgments about those experiences, this would by definition be a fact they had no access to. What's more, since consciousness is an entirely private matter, it would be a fact nobody else had access to either. The absence of any record of such a divergence, then, is entirely unsurprising (in fact it is logically guaranteed), and does not constitute evidence of absence so much as absence of evidence[44].

A reasonable objection to Chalmers' second argument is that the physical changes which occur essentially constantly at the substrate level in humans (e.g. the renewal and replacement of cells and cell components) are of a fundamentally different nature to those described in the fading qualia experiment, and that there is therefore no reason to assume that if the latter introduces a disassociation between experience and experiential judgment the former must as well. Chalmers

[43] Chalmers makes precisely this claim in his own book when he concludes that "what it takes for a state to be a conscious experience in the actual world is for it to have a phenomenal feel, and what it takes for something to be a conscious experience in a counterfactual world is for it to have a phenomenal feel" (1996, p. 118)

[44] This is precisely the same point Cohen and Dennett make in their 2011 paper, when they note that the overflow hypothesis - i.e. the suggestion that we might be phenomenally conscious of more information than we have access to - could never "be tested and examined scientifically" (2011, p. 362).

resists this objection on the grounds that "there seems to be no principled reason why a change from neurons to [simulated neurons] should make a difference while a change in neural realization should not", stating further that "the only place to draw a *principled* line is at the functional level [by which he means the neuronal functional level]" (Ibid., p. 254). This seems like a weak response primarily because there actually do seem to be at least two principled lines that separate the changes in question. Take, for example, the fact that the ordinary cell maintenance and cell replacement processes which occur regularly in the human brain don't radically alter sub-neuronal functioning - cell components are generally replaced by functionally equivalent components, lest the cell stop working properly - while the change described in the fading qualia scenario involves a complete upheaval of such functions (WBE neurons being, as was mentioned earlier, only crude approximations of regular ones that model their input-output function admirably but ignore the bulk of their internal structures and dynamics). A second principled division between the two cases is based on composition: the kinds of substrate-level changes that occur regularly in the human brain involve one biological component being replaced with another essentially chemically identical one; in the fading qualia scenario the replacements that occur are between radically different kinds of substances - carbon-based biological parts on the one hand and virtual structures bottoming out in silicon circuitry on the other. It seems plainly wrong, then, to assert that there are no principled differences between the two cases, so that any dissociation between consciousness and cognition in the one must be mirrored in the other.

There is a stronger challenge which Chalmers could (but doesn't) make to this objection, which is to note that while there may be principled differences between the physical changes involved in the fading qualia hypothetical and those that occur regularly at the substrate-level in humans, the fact that the former can produce a complete dissociation between consciousness and cognition (which was our *reductio* assumption) at least raises the possibility that the latter kind of change might also have a significant but unnoticed effect on our experience. It might be, for example, that the ordinary churn of cell replacement in the brain - though not nearly as drastic as the sort of chemical uprooting Chalmers imagined - might nonetheless result in a kind of shimmering or flickering of our conscious experience that our higher level cognitive processes simply fail to pick up on, and which therefore goes entirely unnoticed. The question then arises whether this is so absurd a possibility as to be able to power the original *reductio*.

Viewed from one angle, the aforementioned possibility seems far from absurd and indeed perfectly reasonable - after all, our cognitive systems evolved to filter out information that it wasn't advantageous for us to pick up on, so we should expect that they would fail to notice changes in our experience (e.g. this hypothetical shimmering) that only carried information about internal cell-replacement processes. Even simple organisms like earthworms seem to do something like this when they distinguish between sensory inputs that are the result of their own movements and those that are due to other entities - something they accomplish by broadcasting so-called 'efference copies' of their own actions around their nervous system which effectively telling their sensors "to ignore some of what comes in: 'don't worry, that's just me'" (Godfrey-Smith 2016, p. 172).

Looked at from another angle, however, the possibility of our experience shimmering in this way, unbeknownst to us, does seem utterly implausible and perhaps even a form of category error. The thinking here is that the very thing we're referring to when we talk about our conscious experience is the content that our mind presents to us after it has filtered our raw sensory inputs through layers upon layers of informational processing. Consider, for example, cases where some object is represented in the visual data that enters through our eye, but fails to be properly identified or categorized by the battery of informational integration processes in our visual cortex, and hence fails to be noticed by us (a real-life example would be the infamous gorilla experiment conducted by Chabris and Simons[45], or - equally powerfully - cases of subliminal visual perception). In these cases, many people think it is more natural to say that the object in question wasn't a part of our visual experience at all, rather than to say that it was there but just unnoticed, or indeed there but unnoticeable (as is the case with subliminal visual perception). If this intuition was correct, then the very idea that substrate-level cellular processes might alter our conscious experiences without our noticing (even in a small way, as with a shimmer or flicker) would be totally nonsensical, since what we're talking about when we use the term experience just is our post-processed content, which we know such cellular processes have no impact on. The fact that the fading qualia scenario raises this possibility, then, would seem a good reason to reject that scenario as absurd, and to accept Chalmers' *reductio*.

This reasoning is compelling, but shouldn't be accepted for the sole reason that it implicitly relies on the very principle that both we and Chalmers dismissed earlier (a principle that, recall, was incompatible with dualism and whose acceptance, therefore, would commit us to a view on the vexed 'what' question of consciousness), namely the principle that noticeability, or else actually being noticed, is a constitutive property of conscious experience. In the absence of such a principle, the possibility outlined above - i.e. that our experiences might be routinely affected (e.g. shimmering) by the substrate-level processes in our brain, and affected in a way that we don't pick up on - ceases to seem absurd, and by extension so does the fading qualia scenario. We arrive, then, at the same conclusion which Shoemaker reached in his response to Chalmers' book[46]: there is one good reason to deny fading qualia scenarios (and so to secure WBE consciousness), but it requires weighing in on the 'what' debate and, more specifically, denying the possibility of dualism.

---

[45] Prinz provides a nice summary of this experiment: "Chabris and Simons (1999) had subjects watch a video in which two teams were tossing a basketball. Subjects were asked to count how many times the ball was passed by a particular team—an attention-demanding task. During the game, a person in a gorilla suit strolls across the center of the screen. The gorilla is highly salient to passive viewers, but 66% of the subjects who were counting passes failed to notice the gorilla" (Prinz 2003, p. 4)

[46] What Shoemaker says specifically is that "anyone who thinks that qualia inversion between functional isomorphs is possible [i.e. any dualist] will also think that qualitative belief inversion between functional isomorphs is possible [i.e. that scenarious like the fading qualia scenario will be possible]" (1999, p. 444)

WHOLE BRAIN EMULATIONS AS MORAL QUESTION-MARKS

Let us recapitulate the positions argued for thus far.

In section I, we argued that - contrary to some environmental theories of moral standing, as well as to certain interpretations of more orthodox moral standing theories (e.g. autonomy, self-awareness, and well-being theories) - there does seem to be a strong link between moral patienthood and conscious status. In particular, we used Carruthers' "Phenumb" thought experiment to demonstrate that a conscious entity's moral status would always be significantly lower if they were non-conscious, holding everything else (i.e. all outward behaviour) equal. We concluded, therefore, that it would be impossible to finally determine an entity's moral status without first determining its conscious status.

In section II, we introduced WBEs and argued that the conscious status of this particular class of AIs could not be ascertained empirically. Our argument began by noting that ascriptions of consciousness to non-human entities are generally based on two separate lines of evidence: analogues of human behavioural correlates of consciousness, and analogues of human neural correlates of consciousness. We observed that the strongest cases for non-human consciousness (e.g. avian consciousness) drew on both forms of evidence, but that inspection of some non-human entities (e.g. fish, but arguably also WBEs) actually revealed a divergence of these two evidential strands, with the entity's behaviour suggesting a capacity for consciousness but its neurology suggesting the opposite. We saw that the only way to determine the conscious status of such entities would be to first determine which kind of evidence - behaviouro-cognitive or neurological - was more fundamental, and that this dispute could be thought of as a disagreement over what David Chalmers called 'bridging principles'. Accepting that such principles draw their justification from their ability to plausibly explain regularities in our conscious data, we considered how the two families of principles in question (behavioro-cognitive and neural) measured up using this metric. We found that, contrary to Chalmers, at least some neural bridging principles (those we termed substrate neural principles) did not compare unfavourably to their behavioro-cognitive counterparts, but rather had unique advantages (namely increased fit and precision) and disadvantages (namely complexity) that at least plausibly balanced each other out. We concluded, therefore, that there actually were - as Chalmers had feared might be the case - two indistinguishably plausible bridging principles, rather than one clear victor. Since one of these principles favoured WBE consciousness and the other ruled it out, we further concluded that the conscious status of WBEs could not be determined empirically.

In section III, we considered David Chalmers' attempt to determine the conscious status of WBEs non-empirically by way of his fading qualia argument. That argument, which took the form of a *reductio*, relied on us finding it implausible that there could be an entity whose cognitive capacities were functioning properly, but which was nonetheless systematically out of touch with its own conscious experience. We considered the three arguments Chalmers raised in

support of this intuition, but ultimately found each unconvincing - the first two facing decisive objections, and the third being at odds with metaphysical dualism. We concluded that any observer not wishing to commit themselves to a view on the vexed question of the ontological nature of consciousness (i.e. the hard problem) would have no reason to accept Chalmers' *reductio* argument, with the result that some other argument would need to be provided if the conscious status of WBEs was to be determined non-empirically.

Tying these sections together, we have the result that - unless some as yet unknown argument can succeed where Chalmers' fading qualia argument failed[47], or some as yet unknown conscious data is discovered[48] which decides the dispute between behavioural and neural bridging principles - there will be no empirical or non-empirical way of determining the conscious status, and by extension the moral status, of WBEs. We arrive, then, at the position foreshadowed in the introduction: that WBEs are moral question-marks whose proper treatment, and whose rightful place in our society, we may simply never know.

[47] Positive arguments for plain physicalist functionalism - arguments which are anyway hard to come by - will not suffice for this purpose, since WBEs are not entirely functionally identical to human brains; they are only functionally identical when considered at a super-neuronal grain of coarseness.
[48] For example Koch and Tononi's Information Integration Theory (2015), which is effectively a higher-level neural bridging principle, makes specific predictions which could be tested (such as at what exact point in the severing of a person's corpus callosum their experience should split in two) and which, if verified, would make their principle seem far more plausible than other competitors.

POSTSCRIPT: A POLICY QUANDARY?

It is tempting to leap from the conclusion defended above (that WBEs are moral question-marks) to the further conclusion that we will never have any good way of factoring such AIs into our moral decision-making processes (a conclusion that would have serious implications for the formulation of equitably social policy). This, however, does not follow without at least some further argument. In particular, there is at least one intuitive proposal for how WBEs might be accounted for in our decision-making (their uncertain moral status notwithstanding) that would have to be examined and rejected before this further conclusion can be plausibly asserted. The proposal is this: why not just give WBEs the benefit of the doubt with respect to their conscious and moral status?

Now there are actually two ways of cashing out this proposal: on the one hand, it could be read as recommending that we assign a one-hundred percent probability to WBEs being conscious, just in case they are (this seems to be what Anders Sandberg's "Principle of assuming the most" (2014, p. 445) is suggesting); on the other, it could be read as simply recommending that we assign WBE consciousness an artificially high probability (e.g. perhaps eighty percent). Something like the latter principle seems to be endorsed by quite a few people in relation to fish-pain (whose conscious status is also, at least at the present, unclear), where it is referred to as the "precautionary principle" (Seth 2016, p. 1).

The first of these two variants of the proposal seems quite obviously misguided, not least because it is out of keeping with accepted approaches to other well-known cases of the same sort of moral uncertainty. Take, for example, our treatment of comatose human patients in the age before brain-scanning technology was advanced enough to reveal whether they still possessed any inner mental life. Such patients were typically given "custodial care" (in which their normal bodily functions were sustained to keep them alive), but were never treated as on a par with other conscious patients, and nor should they have been; in most hospitals, they were deprived of analgesics (which were not in tight supply, and needed by those who were clearly in conscious pain), and were eventually taken off life-support provided no recovery seemed possible (again so that hospital resources could be diverted to definitely-conscious patients). These decisions always carried a risk that they might be inflicting a great harm upon a genuinely conscious being, but the alternative decision carried risks as well, namely the risk that millions of health dollars might be wasted sustaining what was, in effect, a mere shell that had once been - but was no longer - occupied by someone.

The second variant of the 'benefit of the doubt' proposal, such as is commonly advocated with respect to fish, seems much more in line with the historical approach just discussed. One worry about this proposal is that, while it might be appropriate in human cases, applying it to fish results in utterly implausible recommendations. Robert Jones, for example, notes that somewhere between "970 to 2,700 billion fishes are caught from the wild annually" (2016, p. 2), with most of these being killed in quite atrocious ways (e.g. trapping them together in massive

nets and pulling them onto boats to asphyxiate among mounds of their brethren). If this is true, and if we were to take the precautionary principle seriously and artificially assign a probability of something like eighty percent to fish being consciousness, then the commercial fishing industry would appear to be a vast engine of conscious suffering (suffering that, as we argued in section 1, is no trifling ethical matter). It might seem, then, that our moral norms dictate we cease this apparently atrocious practise immediately, economic ramifications notwithstanding. Many take results like this to be obviously absurd, and to act as an effective *reductio* against the precautionary principle itself, but I think this is a weak line of argument that is open to at least two strong objections.

In the first place, the fact that taking a certain ethical principle seriously makes current human practices seem wildly unethical does not seem like a good reason to abandon it, particularly if it seems independently plausible. Had this sort of thinking prevailed in earlier human history, numerous practices that we now - with the benefit of hindsight - consider to have been deeply unethical would have been forever entrenched, e.g. the disenfranchisement of women, and the owning and trading of slaves. In the second place, and perhaps even more importantly, it must be noted that assuming fish are conscious - and capable of experiencing pain consciously - does not imply that fish pain must be as morally significant as human pain. In fact, almost every party to the fish pain debate agrees that, if fish do feel pain consciously, their experience of it would differ radically from our own (being produced, as it is, by a very different and much simpler neural system). One difference which seems of particular importance, and which *prima facie* provides a powerful reason for thinking that fish pain is indeed less morally significant than human pain, is the fact that fish are likely incapable of either meta-cognition or of self-consciousness - that is, fish almost certainly cannot have thoughts about their mental contents (e.g. that this experience is pain) or about themselves (e.g. that it is me that is in pain). Most theorists take this to be one of the most important moral dimensions of negatively valent experiences like pain, with some philosophers (e.g. Tye) even taking them to be the only seriously important one, so the fact that fish pain likely lacks this feature would significantly reduce the extent to which a practise like commercial fishing would have to be (perhaps implausibly) cast as a global atrocity if we took up the precautionary principle.

There is, however, a better reason to resist adopting the precautionary principle when it comes to WBEs, and that is that the central decisions we are likely to have to make concerning WBEs are deeply morally disanalogous to the central decision we have to make concerning fish (namely, whether to keep harvesting them on an industrial scale or not). In the case of fish, there are essentially two possibilities: either fish do feel pain consciously, in which case commercial fishing would seem to constitute a great moral wrong (though exactly how great is unclear), or else they do not feel pain consciously, in which case commercial fishing practices are highly profitable but basically morally neutral (perhaps they are slightly morally praiseworthy insofar as they supply cheap nourishment to certain communities that might struggle to feed themselves otherwise). In a situation like this, adopting something like the precautionary principle seems justified insofar as it guarantees that - at the very least - we will avoid

perpetrating any grave moral crimes. The problem, I argue, is that this is nothing like the decision-problem we face with WBEs.

Whereas using fish for food necessarily requires, at some stage, killing those creatures in vast numbers, WBEs could easily be used without either harming or killing them. Consider, for example, that probably the central application of WBEs (provided they can be manufactured at low enough costs) would be to simply replace human cognitive labour, so that - where a problem might currently be solved by bringing in a human professional such as an engineer or lawyer or doctor - that problem would, in the future, instead be solved by bringing in a simulation of such a professional (e.g. perhaps a WBE based on a world expert in the relevant field). Provided that these WBEs were not turned off between projects (i.e. killed once their work was done), and provided that they be allowed reasonable time to enjoy themselves alongside their work (something they would certainly desire to do, since they are - after all - models of human minds), this sort of application would seem thoroughly morally disanalogous to the industrial-scale slaughter of fish. Indeed, the moral disanalogy becomes even starker when we take account of the fact, which Anders Sandberg brings up, that WBEs could in theory be run at much faster "subjective rates of time" than regular human beings, so that they experienced whole years of their own lives in what - to us - seemed like only a matter of seconds. Given this possibility, it would not be out of the question for WBE workers to be given vast amounts of subjective relaxation time in the objective slivers of time between their various work projects. A WBE doctor might, for example, be permitted to take a subjective month in the Bahamas between one diagnosis and the next, so that the cognitive labour it ended up performing for us over its life would amount to little more than an occasional nuisance.

This is not to say that WBEs will definitely or even probably be applied in such benign and ethically unproblematic ways - it seems quite likely, in fact, that certain profit-interested parties might want to run WBEs like fast-paced simulated slaves, extracting a lifetime of labour from them in a matter of seconds and terminating them once they wear out. Rather, it is to point out that there is nothing inherently morally repugnant about using WBEs to replace human labour. This is important because it means that the central decision to be made concerning WBEs (whether and to what extent to use them) is not - as it is with fish - a choice between one action that is potentially morally atrocious and one that is not (i.e. do or do not allow fish to be commercially harvested), but between a whole range of actions which cover the full spectrum of moral reprehensibility from outrageous to anodyne. Something like the precautionary principle might be useful in paring down this list of options - for example in ruling out those policy suggestions that would be truly morally egregious if WBEs did turn out to be conscious - but beyond that it seems difficult to see how its application could be defended. It would appear, for example, that in considering more narrow policy questions such as what ratio of subjective leisure time to work time should be mandated for WBEs, that simply assigning an eighty-percent probability to WBEs being conscious (and performing some sort of expected-value calculation), would distort the matter in a way that would be difficult to justify.

What has just been said does not amount to, nor was it intended to be, a knock-down argument against the view that we might somehow account for WBEs in our moral decision-making processes despite their uncertain moral status (a full defence of this position would likely require a thesis unto itself, and perhaps more than one). Rather, it was supposed to be a sketch of some compelling reasons to think that this might be so - namely that the only really intuitive proposals for factoring WBEs into our decision-making (i.e. 'benefit of the doubt' proposals) seem to run into serious objections. What does seem clear from the above, and which will suffice for our purposes, is that WBEs raise a uniquely thorny problem for the formulation of morally responsible policy - a quandary that thoughtful politicians of the future are likely one day (and potentially one day soon) to have to seriously grapple with.

ACKNOWLEDGEMENTS

I would like to give particular thanks to Samuel Shpall for his continuous and incisive feedback, as well as his encouragement - both of which he gave in generous helpings.

In addition, I would also like to thank Peter Godfrey-Smith for both his time and invaluable reading suggestions, and Anders Sandberg for his helpful email correspondence.

BIBLIOGRAPHY

(Harvard Referencing Style)

1. Agnieszka, J & Tannenbaum, J 2017, *The Grounds of Moral Status*, Stanford Encyclopedia of Philosophy, 21 September 2017, www.plato.stanford.edu/entries/grounds-moral-status.
2. Akins, K 1993, 'What is it Like to be Boring and Myopic?', in B Dahlbom (ed.), *Dennett and His Critics: Demystifying Mind*, Blackwell Publishing, Cambridge, MA.
3. Allen, C 2004, 'Animal Pain', *Nous*, vol. 38, no. 4, pp. 617-643.
4. Anderson, E 2004, 'Animal Rights and the Values of Nonhuman Life', in C Sunstein & M Nussbaum (eds.), *Animal Rights: Current Debates and New Directions*, Oxford University Press, Oxford.
5. Anderson, M & Anderson, S.L 2011, *Machine Ethics*, Cambridge University Press, Cambridge.
6. Balcombe, J 2016, *Cognitive Evidence of Fish Sentience*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1059&context=animsent.
7. Beukema, J.J 1970, 'Angling experiments with carp', *Netherlands Journal of Zoology*, vol. 20, pp. 81-92.
8. Bloch-Budzier, S 2016, 'NHS using Google technology to treat patients', *BBC News*, 22 November 2016, www.bbc.com/news.
9. Block, N 1978, 'Troubles with Functionalism', *Minnesota Studies in the Philosophy of Science*, vol. 9, pp. 261-325.
10. Block, N 1995, 'On a Confusion About a Function of Consciousness', *Behavioural and Brain Sciences*, vol. 18, pp. 227-287.
11. Block, N 2007, 'Consciousness, accessibility, and the mesh between psychology and neuroscience', *Behavioural and Brain Sciences*, vol. 30, pp. 481-548.
12. Block, N 2011, 'Perceptual Consciousness Overflows Cognitive Access', *Trends in Cognitive Sciences*, vol. 15, no. 12, pp. 567-575.
13. Bosker, B 2017, 'Mayonnaise, Disrupted', *The Atlantic*, November, www.theatlantic.com.
14. Bostrom, N 2014, *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, Oxford.

15. Braithwaite, V.A & Droege, P 2016, *Why Human Pain Can't Tell Us Whether Fish Feel Pain*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1041&context=animsent.

16. Carruthers, P 1999, 'Sympathy and Subjectivity', *Australasian Journal of Philosophy*, vol. 77, no. 4, pp. 465-482.

17. Carruthers, P 2004, 'Suffering Without Subjectivity', *Philosophical Studies*, vol. 121, no. 2, pp. 99-125.

18. Carruthers, P 2005, 'Why the Question of Animal Consciousness Might Not Matter Very Much', *Philosophical Psychology*, vol. 18, no. 1, pp. 83-102.

19. Caruso, C 2017, 'Time to Fold, Humans: Poker-Playing AI Beats Pros at Texas Hold'em', *Scientific American*, 2 March 2017, www.scientificamerican.com.

20. Chalmers, D 1996, *The Conscious Mind: In Search of a Fundamental Theory*, Oxford University Press, Oxford.

21. Chalmers, D 1997a, 'On the Search for the Neural Correlate of Consciousness', in S Hameroff, A Kazniak & A Scott (eds.), *Toward a Science of Consciousness II: The Second Tucson Discussions and Debates*, The MIT Press, Cambridge, MA.

22. Chalmers, D 1997b, *The Problems of Consciousness*, David Chalmers, www.consc.net/papers/montreal.html.

23. Chalmers, D 2010, 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies*, vol. 17, pp. 7-65.

24. Chalmers, D 2014, 'How Do You Explain Consciousness?', March 2014, www.ted.com/talks.

25. Cohen, M.A & Dennett, D.C 2011, 'Consciousness Cannot be Separated From Function', *Trends in Cognitive Sciences*, vol. 15, no. 8, pp. 358-364.

26. Cowey, A & Stoerig, P 1995, 'Blindsight in Monkeys', *Nature*, vol. 373, no. 6511, pp. 247-249.

27. Csanyi, V, Csizmadia, G & Miklosi, A 1989, 'Long-term memory and recognition of another species in the paradise fish', *Animal Behaviour*, vol. 37, no. 6, pp. 908-911.

28. Damasio, A & Damasio, H 2016, *Pain and Other Feelings in Animals*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1064&context=animsent.

29. Dehaene, S & Naccache, L 2001, 'Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework", *Cognition*, vol. 79, no. 1, pp. 1-37.

30. Edelman, D.B, Baars, B.J & Seth, A 2005, 'Identifying Hallmarks of Consciousness in Non-Mammalian Species', *Consciousness and Cognition*, vol. 14, no. 1, pp. 169-187.

31. Edelman, D.B & Seth, A 2009, 'Animal Consciousness: A Synthetic Approach', *Trends in Neurosciences*, vol. 32, no. 9, pp. 476-484.

32. Edelman, D.B 2016, *Leaving the Door Open for Fish Pain: Evolutionary Convergence and the Utility of 'Just-So Stories'*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1066&context=animsent.

33. Elliot, R 1997, *Faking Nature: The Ethics of Environmental Restoration*, Routlege, London.

34. Elwood, R.W 2016, *A Single Strand of Argument with Unfounded Conclusion*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1056&context=animsent.

35. Farah, M 2008, 'Neuroethics and the Problem of Other Minds: Implications of Neuroscience for the Moral Status of Brain-Damaged Patients and Nonhuman Animals', *Neuroethics*, vol. 1, no. 1, pp. 9-18.

36. Feldman, F 2010, *What is This Thing Called Happiness*, Oxford University Press, Oxford.

37. Gamez, D 2008, 'Progress in Machine Consciousness', *Consciousness and Cognition*, vol. 17, no. 3, pp. 887-910.

38. Gibbs, S 2017, 'Elon Musk leads 116 experts calling for outright ban of killer robots', *The Guardian*, 21 August 2017, www.theguardian.com.

39. Gibbs, S 2017, 'Elon Musk: Tesla electric lorry to be unveiled in late October', *The Guardian*, 14 September 2017, www.theguardian.com.

40. Godfrey-Smith, P 2016, *Pain in Parallel*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1057&context=animsent.

41. Godfrey-Smith, P 2016, *Other Minds: The Octopus, The Sea, and the Deep Origins of Consciousness*, Farrar, Straus, and Giroux Publishing, NY.

42. Godfrey-Smith, P 2017, 'The Evolution of Consciousness in Phylogenetic Context', in K Andrews & J Beck (eds.), *The Routlege Handbook of Philosophy of Animal Minds*, Routlege, London.

43. Good, I.J 1966, 'Speculations Concerning the First Ultraintelligent Machine', *Advances in Computers*, vol. 6, pp. 31-88.

44. Jones, R.C 2016, 'Fish Sentience and the Precautionary Principle', Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1032&context=animsent.

45. Kant, I 1998, *Groundwork of the Metaphysics of Morals*, trans. M Gregor, Cambridge University Press, Cambridge, original work published 1785.

46. Key, B 2015, 'Fish do not feel pain and its implications for understanding phenomenal consciousness', *Biology and Philosophy*, vol. 30, no. 2, pp. 149-165.

47. Key, B 2016a, *Why Fish Do Not Feel Pain*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1011&context=animsent.

48. Key, B 2016b, *Going Beyond Just-So Stories*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1045&context=animsent.

49. Khatchadourian, R 2015, 'The Doomsday Invention', *The New Yorker*, 23 November 2015, www.newyorker.com.

50. Korsgaard, C 1992, 'The Sources of Normativity', 16-17 November 1992, www.dash.harvard.edu.

51. Korsgaard, C 2004, 'Fellow Creatures: Kantian Ethics and Our Duties to Animals', 6 February 2004, www.dash.harvard.edu.

52. Korsgaard, C 2007, 'Facing the Animal You See in the Mirror', 24 April 2007, www.people.fas.harvard.edu.

53. Levine, J 1983, 'Materialism and Qualia: The Explanatory Gap', *Pacific Philosophical Quarterly*, vol. 64, pp. 354-361.

54. Lin, P 2013, 'The Ethics of Autonomous Cars', *The Atlantic*, 8 October 2013, www.theatlantic.com.

55. Lori, G 2017, *The Moral Status of Animals*, Stanford Encyclopedia of Philosophy, 21 September 2017, www.plato.stanford.edu/entries/moral-animal/.

56. Mather, J.A 2016, *An Invertebrate Perspective on Pain*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1046&context=animsent.

57. Millsopp, S & Laming, P 2008, 'Trade-offs Between Feeding and Shock Avoidance in Goldfish', *Applied Animal Behaviour Science*, vol. 113, pp. 247-254.

58. Naccache, L, Blandin, E & Dehaene, S 2002, 'Unconscious Masked Priming Depends on Temporal Attention', *Psychological Science*, vol. 13, no. 5, pp. 416-424.

59. Nagel, T 1974, 'What is it Like to Be a Bat?', *The Philosophical Review*, vol. 83, no. 4, pp. 435-450.

60. Ohnsman, A 2017, 'At $1.1 Billion Google's Self-Driving Car Moonshot Looks Like A Bargain', *Forbes*, 15 September 2017, www.forbes.com.

61. Prinz, J 2003, 'A Neurofunctional Theory of Consciousness', in A Brook & K Akins (eds.), *Philosophy and Neuroscience*, Cambridge University Press, Cambridge.

62. Prinz, J 2012, *The Conscious Brain*, Oxford University Press, Oxford.

63. Prinz, J 2017, 'Attention, Working Memory, and Animal Consciousness', in K Andrews & J Beck (eds.), *The Routlege Handbook of Philosophy of Animal Minds*, Routlege, London

64. Quinn, W 1984, 'Abortion: Identity and Loss', *Philosophy and Public Affairs*, vol. 13, no. 1, pp. 24-54.

65. Regan, T 1985, 'The Case for Animal Rights', in P Singer (ed.), *In Defence of Animals*, Blackwell Publishing, Oxford.

66. Robbins, M 2016, 'Has a rampaging AI algorithm really killed thousands in Pakistan?', *The Guardian*, 19 February 2016, www.theguardian.com.

67. Rose, J.D 2002, 'The Neurobehavioral Nature of Fishes and the Question of Awareness and Pain', *Reviews in Fisheries Science*, vol. 10, no. 1, pp. 1-38.

68. Rose, J.D 2014, 'Can Fish Really Feel Pain?', *Fish and Fisheries*, vol. 15, no. 1, pp. 97-133.

69. Sandberg, A 2014, 'Ethics of Brain Emulations', *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 26, no. 3, pp. 439-457.

70. Searle, J 1990, 'Who is Computing With the Brain?', *Behavioural and Brain Sciences*, vol. 13, no. 4, pp. 632-642.

71. Seth, A, Baars, B.J, & Edelman, D.B 2005, 'Criteria for Consciousness in Humans and Other Mammals', *Consciousness and Cognition*, vol. 14, no. 1, pp. 119-139.

72. Seth, A 2010, *The Grand Challenge of Consciousness*, Frontiers in Psychology, www.ncbi.nlm.nih.gov/pmc/articles/PMC3153737/.

73. Seth, A 2016, *Why Fish Pain Cannot and Should Not Be Ruled Out*, Animal Sentience, animalstudiesrepository.org/cgi/viewcontent.cgi?article=1038&context=animsent.

74. Shead, S 2017, 'Google DeepMind's AlphaGo AI beat the best Go player in the world in its first game', *Business Insider*, 23 May 2017, www.businessinsider.com.au.

75. Shoemaker, S 1999, 'Review: On David Chalmers's the Conscious Mind', *Philosophy and Phenomenological Research*, vol. 59, no. 2, pp. 439-444.

76. Singer, P 1993, *Practical Ethics*, Cambridge University Press, Cambridge.

77. Sneddon, L.U, Braithwaite, V.A, & Gentle, M.J 2003, 'Do Fishes Have Nociceptors? Evidence for the Evolution of a Vertebrate Sensory System', *Proceedings: Biological Sciences*, vol. 270, no. 1520, pp. 1115-1121.

78. Tononi, G 2008, 'Consciousness as Integrated Information: a Provisional Manifesto', *The Biological Bulletin*, vol. 215, no. 3, pp. 216-242.

79. Tononi, G & Koch, C 2015, 'Consciousness: Here, There, and Everywhere?', *Philosophical Transactions of the Royal Society of London*, vol. 370, issue. 1668.

80. Tooley, M 1972, 'Abortion and Infanticide', Philosophy and Public Affairs, vol. 2, pp. 37-65.

81. Tye, M 2000, *Consciousness, Colour, and Content*, The MIT Press, Cambridge, MA.

82. Urban, T 2015, 'The AI Revolution: The Road to Superintelligence', *Wait But Why*, 22 January 2015, www.waitbutwhy.com.

83. Van Gulick, R 2014, *Consciousness*, Stanford Encyclopedia of Philosophy, 21 March 2014, www.plato.stanford.edu/entries/consciousness/.