The hard problem of AI rights

Adam J. Andreotta¹

Received: 14 February 2020 / Accepted: 19 May 2020 / Published online: 30 May 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract

In the past few years, the subject of AI rights—the thesis that AIs, robots, and other artefacts (hereafter, simply ‘AIs’) ought to be included in the sphere of moral concern—has started to receive serious attention from scholars. In this paper, I argue that the AI rights research program is beset by an epistemic problem that threatens to impede its progress—namely, a lack of a solution to the ‘Hard Problem’ of consciousness: the problem of explaining why certain brain states give rise to experience. To motivate this claim, I consider three ways in which to ground AI rights—namely: superintelligence, empathy, and a capacity for consciousness. I argue that appeals to superintelligence and empathy are problematic, and that consciousness should be our central focus, as in the case of animal rights. However, I also argue that AI rights is disanalogous from animal rights in an important respect: animal rights can proceed without a solution to the ‘Hard Problem’ of consciousness. Not so with AI rights, I argue. There we cannot make the same kinds of assumptions that we do about animal consciousness, since we still do not understand *why* brain states give rise to conscious mental states in humans.

Keywords Artificial intelligence · Ethics · AI rights · Animals rights · The ‘hard problem’ of consciousness

1 Introduction

Artificially intelligent robots, machines, and other artefacts (hereafter, simply ‘AIs’) already reside in our workplaces—from unmanned drones, to warehouse packing robots; as well as in our homes—from social robots such as ‘Jibo’, to virtual assistants such as Amazon’s Echo/Alexa. Some (see, e.g., Bostrom 2014; Tegmark 2017; Russell 2019) worry about what will happen when future AIs surpass human level intelligence. Will we eventually lose the ability to control them? Nick Bostrom, for example, claims that the prospect of superintelligent AIs ‘is quite possibly the most important and most daunting challenge humanity has ever faced’ (2014, p. v). According to others (see, e.g., Putnam 1964; Wallach and Allen 2009; Coeckelbergh 2012; Schwitzgebel and Garza 2015; Gunkel 2012, 2018) we should be concerned about the moral standing of AIs. Such authors have focused on the question of whether AIs should ever be endowed with legal or moral rights; just like human beings, certain nonhuman animals, and the environment are today.

In this paper, I examine the latter issue. I consider the two-part question posed by David Gunkel: ‘Can and should robots have rights?’ (2018, p. 2). Adopting a broad conception of AI that includes robots, machines and other artefacts, I argue that AIs can and should have rights—but only if they have the capacity for consciousness. This mirrors the same reasoning that is commonly employed in discussions about animal rights. Problematically, for AI rights, however, the analogy with animal rights is not perfect. Since we share an evolutionary history with mammals, birds, reptiles, and other nonhuman animals (hereafter, simply ‘animals’), we are entitled to make certain assumptions about the experiences they undergo based on our common biology. One drawback of this method is that the less an animal resembles a human, the harder it is to know the extent to which its experience resembles our own, or if it even has experiences at all. Given that advanced AIs will likely be constituted in ways that are very different to us, I argue that current approaches to animal consciousness do not map well to questions of AI consciousness. The ‘Hard Problem’ for AI rights, I contend, stems from the fact that we still lack a solution to the ‘Hard Problem’ of consciousness—the problem, as David Chalmers puts it, of why certain functions or brain states are ‘*accompanied by experience*’ (2010, p. 8 emphasis in original). Why for example do some brain

✉ Adam J. Andreotta
adamandreotta@outlook.com

¹ Department of Philosophy, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia

states give rise to red experiences and others green, others pain experiences, and others nothing at all (e.g., when one is under general anaesthesia)? When it comes to the problem of animal consciousness, we can sidestep such questions while still making progress on the problem, given our close biological resemblance to certain animals. Not so with AI consciousness I argue. Since AIs are and will continue to be constituted in ways that differ greatly from us—at least for the foreseeable future—we will not be able to circumvent the ‘Hard Problem’ if we wish to address the question of AI rights. This is a pessimistic conclusion, given the ‘Hard Problem’ eludes a widely agreed upon solution. However, it does have two main upshots. The first is that the ‘Hard Problem’ is important to solve for practical reasons—namely, to address the question of AI rights; the second is that those who have characterised the ‘Hard Problem’ as a pseudo problem, a kind of ‘illusion’ (see Frankish 2016), or a ‘distraction’ (see Dennett 2018, p. 1) have done so prematurely.

I proceed as follows. In Sect. 2, I provide a general overview of rights and distinguish direct rights/duties from indirect rights/duties. I then show why this distinction is relevant to the problem of AI rights. In Sect. 3, I argue that superintelligence in and of itself is not enough to ground AI rights. In Sect. 4, I argue that empathy is a problematic way to ground AI rights. In Sect. 5, I explain why consciousness is required to establish direct rights in general. In Sect. 6, I show how the question of animal consciousness, and thus animal rights, can circumvent the ‘Hard Problem’ of consciousness. In Sect. 7, I show how current approaches to determining whether AIs are conscious which sidestep the ‘Hard Problem’ are unsatisfactory.

2 On AI rights

Before addressing the question of whether rights ought to be extended to AIs, it is important that the concept of rights is first explicated.¹ While all of us make right assertions from time to time, and thus have some familiarity with the concept, the term ‘rights’ lends itself to multiple senses that are important to distinguish (Gunkel 2018, p. 26). What is a right, and what if anything is the essential property that all rights possess? While I cannot fully account for the rich rights literature that addresses this question here, it will be useful to mention Wesley Hohfeld’s influential account of the internal structure of rights. These are known in the

literature as Hohfeld’s (or Hohfeldian) Incidents (Wenar 2005) and will help to set the stage for the discussion on AI rights to come.

Hohfeld (1919) offered four different characterisations of rights, while originally for legal rights, they are also applicable to moral, political, and cultural rights (Gunkel 2018, p. 27). First, there are *privileges*: these include acts for which one has no legal (or moral) duty not to perform. An ambulance driver (unlike the average person driving to work), for example, has a right to run a red light, because she has no legal (or moral) duty to stop. Second there are *claims*: these include right assertions that give rise to positive duties in other people. If I have a right to the 200 dollars that a friend borrowed from me, for example, then that friend has a duty to pay the money back. The third of the Hohfeldian incidents is *power*: these are rights that come about from authority (Wenar 2005). A boss has a *right* to fire an employee if she catches the employee stealing from her company. This power is not given to all employees and is of course limited: the boss does not have the right to fire an employee on racist or sexist grounds. Fourth, there are *immunities*. These kinds of rights arise in situations, where a person (or animal) has a right not to be interfered with or harmed (Wenar 2005, p. 232). As with rights that are claims, immunities give rise to duties in other people. For example, if a family of Gorillas is being hunted by poachers, then the local government has a duty to protect the Gorillas by arresting the poachers. The poachers would here be ignoring their legal (as well as moral) duty to avoid hunting the Gorillas, and thus would lose their right to immunity.² Finally, these categories are not mutually exclusive: some rights can be characterised in terms of more than one category (Wenar 2005, p. 229).³

When rights work as they ought to, the result is a fairer and more just society; one that allows individuals the freedom to pursue their own interests. In addressing the question of whether AIs should have rights, I will mainly be concerned with two Hohfeldian incidents: *duties* (e.g., Humans have a duty not to put AIs in danger) and *immunities* (e.g., Humans ought not to enslave AIs). I will be less concerned with privileges and powers that AIs ought, or ought not, to be granted, as I think this will depend upon what kinds of AIs we build. The argument I am making here is a general one. Towards this end, a further distinction is crucial to make, one that is not always made in the literature—namely, the one between direct/indirect rights and direct/indirect

¹ In addressing the question of what a right is, and the further problem of AI rights, I am operating within a Western Philosophical tradition of rights, laws, and normative ethics. I acknowledge that the concept of ‘rights’ is understood differently in Eastern and First Nations/Indigenous traditions.

² According to some, such as Kramer (1998, p. 9), all rights claims have duties associated with them.

³ There are two main theories on the proper function of rights in the literature—namely: the will theory, and the interest theory. The argument I am making in this paper does not presume the truth of either. See Kramer (1998) for a debate on this issue.

duties. A direct rightsholder can be defined as an entity (e.g., a human, or animal) who is owed a certain state of affairs (e.g., freedom, protection, education) by a person (or group of people) who has duties to that entity (e.g., to protect them). An indirect rightsholder by contrast, can be defined as an entity (e.g., a work of art, a car) from which rights and duties arise from, but is not owed anything directly. The duties that we have to these entities (e.g., to protect them) are indirect, meaning that they only come about because of the duties that are owed to someone else.

To illustrate this distinction, consider my duty not to steal another person's car. Here I must respect the owner's claim—that is, I have a duty not to steal her car, given that I have a duty to respect her property rights. Now, while this right *arises* from the car, my *duty* is clearly not to the car—it is to the owner. Kant famously employs this distinction in his discussion of our duties to animals (Howe 2019). Kant thinks we should refrain from hurting animals because of how it might cultivate our own character—not because of what it does to them. He states, 'he who is cruel to animals becomes hard also in his dealings with men' (1963, p. 240). Here we can see that Kant does not think we have any direct duties to animals. Following Robert Garner (2013), let us call this the *indirect rights/duties* view. There are well known difficulties with the indirect rights/duties view when applied to animals, however. The reason factory farming is morally problematic, for example, is precisely because of the harm it causes to animals, not because of how it affects us. Bestowing only indirect rights, and corresponding duties, to animals does not take into consideration their suffering. So, if we accept the indirect rights view, with respect to animal rights, then the horrors of factory farming would be morally justified, as long as it does not negatively impact humans. Given such an implication, Garner correctly notes that 'it is doubtful if a credible indirect duty approach to animals can be developed' (2013, p. 9).

Few philosophers today would say, along with Kant (and also Descartes), that we only have indirect duties to animals. There are some, however, that do. Carruthers (1992) is a proponent of the indirect rights/duties view, with respect to animals. He claims, 'I am morally obliged not to kill your dog, just as I am obliged not to set light to your car' (1992, p. 106). Carruthers notes that if he were to set light to your dog it would be '*your* rights that I would infringe, not the dog's. Indeed, the dog would have no rights, any more than the car does' (1992, p. 106 emphasis in original). For Carruthers, and others who hold this view, there are still limits to how we should treat animals. But these limits, and subsequent duties, are not grounded in animal suffering. They are grounded in the duties we have to the owners of animals—e.g., their property rights. That is why they are indirect.

I do not mean to suggest here that indirect rights are always inappropriate, or inferior, however. In other cases,

they are fitting. Consider the case of New Zealand's Whanganui River, which was recently given 'human rights' and 'personhood', meaning that it can be represented in the courts by a guardian—in this case two people: one appointed by the local Māori people, the other appointed by the New Zealand Government (O'Donnell and Talbot-Jones 2017). While establishing legal frameworks for rivers, such as the Whanganui, is a way to ensure that important steps can be taken to protect them, this does not mean we have direct moral duties to rivers. A river, after all, will not have its desires frustrated, interests ignored, or suffer in a conscious sense, if it is destroyed. Rather it is we humans, and other creatures who rely on it to survive, who suffer if the river is destroyed. Giving the Whanganui River rights (albeit indirect ones) gives us a formal way of recognising the local people's land and ensures that the river is preserved for future generations. Legal rights and duties, after all, have the benefit of being enforceable. Moral rights, and corresponding duties, on the other hand are often hard to enforce (Kramer 1998, p. 9).

Consider, further, the case of climate change: there we have a duty to stop global warming because of the direct duties we have to future generations of people and animals: we have a moral duty to leave the planet in a liveable condition. However, the climate itself, or planet, does not suffer if we fail at this task. There may even be species on Earth that would flourish on a warmer planet, even if humans were wiped out. Our direct duties, then, are to other humans and certain animal species. None of this is to downplay the severity of the accelerating global crisis. A 'rights revolution in nature' as Chapron et al. (2019, p. 1392) call for, will be necessary to implement, if we wish to stop the destruction of the natural world. Furthermore, Earth Jurisprudence proponents (e.g., Burdon 2012; Rühls and Jones 2016) are right that many existing laws reflect an outdated worldview: they permit the exploitation of the planet for our own good. The ongoing climate change crisis and the 2019–2020 COVID-19 pandemic provide examples of what happens when we get things wrong. Establishing rights for nature helps with the task of reforming existing laws and policies, to achieve such ends. These rights may be indirect, but this does not make them any less important to uphold.

The question to be addressed here then is: 'Should we give direct rights to AIs, or, at best, would we only have indirect duties to them—the kind that we might grant to someone's house or car?' Would protecting them from harm be justified only if it benefitted us; or would the suffering of AIs be bad in itself? Critics of AI rights such as Joanna Bryson, are sceptical that we have (or would have) direct duties to AIs. Her provocatively titled essay 'Robots Should Be Slaves' argues that 'In humanising [AIs] we not only further dehumanise real people, but also encourage poor human decision making in the allocation of resources and

responsibility’ (2010, p. 63). Bryson’s claim only has merit, however, if we really do not have any direct duties to AIs. If one argued that it wastes resources to care about the welfare of animals, one could fairly reply by saying that the allocation of resources is morally justified by pointing out that doing so prevents animal suffering. If AIs really do have the moral status of a toaster, Bryson certainly has a point. But if AIs do have some moral status, then she is mistaken.

What properties would an AI be required to possess, in order for them to be granted moral status, and be given direct rights? In posing the question in this way, I am adopting a position that Mark Coeckelbergh calls the ‘property-based account’ (2012, p. 23; cf. Gunkel 2018, p. 92). According to this view, what gives an entity moral standing depends upon the properties it possesses. Coeckelbergh (2012) himself argues against the property-based account, because he thinks that it would be hard to know which properties would be the morally salient ones, or whether a certain entity had those properties. I agree that these are problems; but I do not take them to undermine the property-based account. Rather, I think they simply present us with a challenge. I begin by considering intelligence as one of these properties.

3 Intelligence

One possibility is that we would have direct duties to AIs that are sufficiently intelligent or complex (see, e.g., Goertz 2002). Such a position may be motivated by observing that we endow greater rights to species with high levels of intelligence or complexity. The great apes, for example, are given substantive rights, whereas sea slugs are not. The question of how intelligent or complex AIs would need to become to warrant such concern is a difficult one. One idea is that AIs would need to have intelligence that comes close to, or reaches, human level intelligence—referred to as ‘artificial general intelligence’ (AGI); or surpasses it—referred to as ‘superintelligence’ (see Bostrom 2014, p. 63). As I will consider consciousness as a condition for AI rights separately in Sect. 7, I shall focus here on the specific question ‘Do AIs which possess AGI or superintelligence deserve moral concern, regardless of whether they are conscious?’ Addressing this circumscribed question will allow us to isolate the extent to which intelligence alone is sufficient for AI rights.⁴ The AIs I consider in this section should be thought of as

philosophical zombies—meaning that they possess no phenomenological conscious experience.⁵

We know what human level intelligence is like, but what would a superintelligent AI look like? Bostrom describes three categories in which AIs may surpass human level intelligence: speed, collectiveness, and quality.⁶ For example, AIs may be able to perform tasks *faster* than us, such as having the capacity to read Tolstoy’s *War and Peace* in under a minute; a collection of individual AIs could work in confluence to perform a complex task, such as terraforming Mars; and AIs could produce greater *quality* output compared to humans, such as coming up with solutions to unsolved problems in physics and mathematics in a matter of seconds. Some AIs which reached this level of complexity, depending on how they were constituted, would also possess a capacity for autonomy (e.g., be able to move around without explicitly being told what to do) and be able to pass the Turing Test—Alan Turing’s (1950) famous test for determining whether machines can ‘think’.⁷

While we would find such abilities—at least from our current perspective—impressive, we need to ask why such technological advances should count ethically. If robot vacuum cleaners such as *Roomba* or chess-playing computers such as *Deep Blue* do not have direct rights, then why should a superintelligent humanoid that can communicate with us, or move without being explicitly told to, deserve rights? One reason is that the former artefacts do not express any preferential (or choice) behaviour: robot vacuum cleaners and chess-playing computers do not act as if they care about whether they are unplugged or are made to work continuously. A superintelligent humanoid AI, on the other hand, could express its preferences about how it was being treated. For example, a superintelligent AI could object to being made to work continuously; express distaste about being subject to experimental testing; or appear agitated when told it will be ‘turned off’.

Since preferential (or choice) behaviour plays an important role in discussions of animal rights, it is not implausible to think that preferential (or choice) behaviour in AIs could help inform our discussions about AI rights.⁸ Questions

⁴ I say ‘sufficient’ here because we would not want to say that high levels of intelligence are necessary for direct rights. Animals, small children, and the mentally impaired, are deserving of such rights and do not meet these high standards.

⁵ The concept of a philosophical zombie is elaborated further by Chalmers (2010, pp. 106–107).

⁶ Bostrom does not claim that superintelligence is sufficient for AIs rights.

⁷ Turing famously thinks that this is a practical definition of thinking. He thinks that the more abstract question ‘Can machines think?’ is ‘too meaningless to deserve discussion’ (1950, p. 442).

⁸ See Gunkel (2012 pp. 42–44) for a discussion about the parallels between animal rights and AI rights. Gunkel notes that the idea of ‘personhood’ has been invoked by animal rights activists in order to give animals legal rights; and notes that, while no one has claimed personhood for any robot yet, there are cases in robot science fiction where personhood for robots has been sought for this same reason (2012, pp. 43–44).

about the ethical treatment of fish, for example, have been informed by experiments on how different species react to various stimuli: fish appear wary and initially avoid novel objects (e.g., a Lego tower) that are placed in their environment (see Braithwaite 2010, pp. 68–69). And the recognition of the complexity of Octopus behaviour has led to the species' legal protection. In 1986, for example, the British Government revised its Animals (Scientific Procedures) Act by adding a clause about not operating on sedated octopuses (see Dennett 1996, p. 74).

This approach when applied to AIs is complicated by the fact that we cannot simply assume an AI's choice behaviours represent 'conscious' preferences. The reason why the fish experiments are pertinent to questions of fish rights is that they may provide support to the claim that fish have actual experiences. One interpretation of the avoidance behaviour observed in fish is that it shows that they may undergo the negative experiences that are associated with pain (Braithwaite 2010, p. 69). If it turned out that this hypothesis was mistaken, however, then this result would have serious implications for the question of fish rights. Similarly with AIs I argue. If an incredibly sophisticated AI communicated a preference not to be destroyed, or protested that its rights were being infringed upon, and did not actually undergo any experiences, then such utterances would not be representing any internally felt experiences. Their utterances would be analogous to the one a robot vacuum cleaner makes when its batteries are low—namely, 'Please charge Roomba'. Such a sound does not represent a conscious preference not to be harmed. It is merely an audio file uploaded by a programmer, to inform users about the artefact's status. If these sounds do not have moral significance, then I do not think that a system which has the capacity to play complex audio files in response to human questioning is automatically deserving of rights.

The idea that an AI's preference, or avoidance, behaviours, may occur without any felt experiences must be taken as a real possibility given the presence of such cases in animals. Rats, for example, who have had their spinal cord split—meaning that their brains cannot register tissue damage, and thus it is unlikely that they feel the pain—can still exhibit 'pain behaviour' and still show a form of learning that relates to the location of the damage (Godfrey-Smith 2016, pp. 93–94). This supports the claim that complex behaviour can occur without the presence of pleasure or pain—the key concern for many when thinking about the rights of animals (c.f. Bentham [1789] 2005; Singer [1975] 2009). As John Rawls puts it, '[t]he capacity for feelings of pleasure and pain and for the forms of life of which animals are capable clearly imposes duties of compassion and humanity in their case.' ([1971] 1999, p. 512). A zombie AI, who cannot experience pleasure or pain, but can avow a desire or preference would merely be *imitating* a conscious

creature. Superintelligent AIs of this variety will only be bearers of indirect rights, however. Rights may arise from them, but we will not have duties towards them. For example, we may wish to ensure they are protected from harm, because doing so would allow us to take advantage of their labour. Our moral obligations to them, however, would be no greater than the ones owed to an expensive piece of machinery.

It is unlikely that humans would, at least initially, react to the superintelligent AIs in the same way they react to kitchen appliances, even if superintelligent AIs lack the capacity for experience. The presence of communication, agency, and resemblance to us would likely invoke a feeling of empathy within us—analogueous to the way that we do with other people and some animals. I now turn to empathy, to assess its moral significance.

4 Empathy

Empathy can be defined as the act of putting oneself in the shoes of another. Jesse Prinz, more formally, describes this as 'a matter of feeling an emotion that we take another person to have' (2011, p. 215). This process involves perspective taking: a mirroring of the emotions or feelings that we take another to be experiencing. Why think that empathy can provide us with a way to ground AI rights? According to some philosophers, such as David Hume, Adam Smith, and more recently Michael Slote (2010), it is our ability to take the perspective of another that gives rise to, or is the foundation of, our moral decisions. Second, some such as Simon-Baron Cohen (2011) explain human cruelty by appealing to a lack of empathy. Thirdly, empathy is often invoked by public figures, such as Barack Obama, who claim that more empathy would create a better society (cf. Bloom 2016, p. 18).

There is much to say about empathy and its role in moral decision making—both with respect to human–human and human–animal interactions. I will limit my attention here to the 'dark side of empathy' that Prinz (2011) and Bloom (2016) have recently explored. They raise a number of objections to the thesis that empathy is always a good guide to compassion. Drawing upon their criticisms, I argue that these problems make empathy an unsatisfactory way to ground AI rights. I consider two problems to support this claim. First, *the similarity bias problem*; and second, *the framing problem*.

The similarity bias problem for empathy arises from the empirical observation that we feel greater empathy for those who are similar to us (cf. Bloom 2016, pp. 93–100; Prinz 2011). For example, Prinz cites empirical data from Xu et al. (2009), who employed brain imaging technology (fMRI) to show that Caucasians were able to empathise with the pain

of other Caucasians to a greater extent than to which they could empathise with Chinese participants; and vice versa. The experimenters concluded that ‘Our fMRI results support the view that shared common membership enhances a perceiver’s empathic concerns for others.’ (2009, p. 8528). Given that morality makes demands upon us that are applicable to people quite different from us, such findings are problematic for the claim that empathy always results in fairness.

These biases in empathetic perspective taking also arise in human–animal interactions: the less a species resembles our own, the less likely we will be able to accurately feel their pain or experiences. For example, it is well known that we are more likely to feel moral concern for ‘cute’ animals, such as baby kittens, or puppies. This may be because they bring out our parental urges (Herzog 2010, p. 39). But it is hard to see how being cute could count morally, however: surely other animals deserve our concern, too. These biases can also lead us to make mistakes about the feelings or emotions we take an animal to have. For example, many people think that since humans appreciate hugs, dogs must too. If one were to put oneself in the position of a dog, they might conclude that a hug would be a pleasurable experience. The trouble is, however, as dog experts tell us, dogs do not like this: it often causes them discomfort (Bloom 2016, p. 66).

If biases in empathetic perspective taking can affect moral reasoning with respect to human–human and human–animal interactions, then moral reasoning with AIs will surely be prone to the same problems. We will presumably find it easier to empathise with AIs that look and act just like us. And find it much harder to empathise with AIs which are very different from us yet still worthy of moral concern. Furthermore, we may even attempt to take the perspective of a zombie AI, and thus project emotions on an AI that has no inner life.

The second problem with grounding AI rights in empathy is the framing problem—this is the problem that empathy can easily be manipulated by the way in which one attempts to take the perspective of another. While there exist data showing this occurring in human–human moral reasoning (see C. Daniel Batson et al. 1995), there are some data to suggest that the same problem will be present with AI moral reasoning. Kate Darling (2017), for example, describes experiments she carried out, where participants were asked to observe, and then hit with a mallet, a Hexbug Nano—a small robotic toy automaton. In the control group, Darling introduced an anthropomorphic framing effect by giving backstories about the Hexbug. One of these was: ‘This is Frank. He’s lived at the Lab for a few months now. His favorite color is red’ (2017, p. 181). Darling found that there was greater hesitation to hit the bug in the group who were given a backstory about the Hexbug. Darling notes that subjects asked questions like ‘Will it hurt him?’, and muttered

under their breath while hitting the Hexbug, ‘It’s just a bug, it’s just a bug’ (2017, p. 181). This shows how easy it is to induce moral concern in creatures who do not suffer at all: it is extremely implausible that the tiny robotic toy automatons undergo any emotions or experiences.

The central problem with grounding AI rights in empathy, then, is that empathy can easily push one’s moral decisions in the direction of injustice. It is not only harder for us to empathise with people who are different from us, but it is easy to make mistakes when we attempt to put ourselves in the shoes of another. For example, if a zombie AI robot looked and acted like a human, and it was given a back story, it would be very natural for us to have sympathy for it if it was deprived of rights. I now turn to consciousness, which I argue provides the best way to ground AI rights.

5 Consciousness as a necessary condition for AI rights

The rights movements that have occurred throughout the past two centuries have led to the illegality of slavery (though it persists); civil rights being granted to African Americans in the 1960s; women being allowed to vote in elections around the world; and the introduction of laws governing how we can treat certain animals. The establishment of these rights, while not always guaranteeing justice, has provided a framework for reducing the suffering of the members of these groups: individuals are better off experientially speaking than they would have been if those rights had not been granted. A caged chimpanzee, for example, who is subject to experimental testing is worse off experientially, *ceteris paribus*, compared to his counterpart who is free to live in his natural habitat.

AI rights—if they are to be established in the direct sense I have in mind—should be built upon the same principle. The aim should be to improve the experiences of AIs—meaning that unnecessary suffering of AIs should be eliminated. It is for this reason that a capacity for consciousness or experience is necessary. (I follow Koch in using the terms ‘experience’ and ‘consciousness’ ‘interchangeably’ (2019, p. 1)). A mobile phone for example, however, poorly treated, does not suffer, because (presumably) it is not conscious. Thus, it would not make sense to minimise its suffering by giving it direct rights, because it cannot have experiences in the first place.

Before getting to the epistemic problem of knowing whether a certain AI is conscious, I will first consider an objection to the view that AI rights can be grounded in consciousness from Gunkel. He claims, we do not ‘have any widely accepted characterization of “consciousness”’ (2014, p. 116); and that there is ‘little or no agreement when it comes to defining and characterizing’ (2014, p. 116)

consciousness. He acknowledges that while consciousness is commonly thought to be more scientific than the old idea of the soul, it is ‘just as much an occult property’ (2018, p. 99). Now, while Gunkel is right that there are many controversies surrounding the nature of consciousness, his comments overlook the fact that conscious experiences, from the first-person perspective, are intimately known to all of us.⁹ Consider being conscious of an agonising pain in your wrist. Such an experience has a phenomenological subjective quality to it, which you can effortlessly comprehend. We can further clarify the concept by imagining what it would be like to lose consciousness of something. It is when pain goes away, for example, that we are no longer aware or *conscious* of the pain (cf. Prinz 2012, pp. 4–5). Similarly, the simple act of closing one’s eyes takes away the conscious experience of one’s visual field and opening them brings one’s conscious experiences back. The term ‘consciousness’ may be ‘multiply ambiguous’ as Prinz (2012, p. 4) put it, but there is a sense in which it is easily comprehensible—namely, when we are talking about first-person experience.

The problem, then, with grounding AI rights in consciousness is not, as Gunkel claims, that it is a mysterious ‘occult’ property. The problem is an epistemological one—that is, how we can know whether consciousness exists in other creatures, and furthermore, how we know what those experiences are like. Few today would be as restrictive as Descartes, who famously thought that animals were like machines who lacked consciousness—that is, they lacked sensations and emotions. Most would grant that our primate relatives, such as the great apes and monkeys, have experiences that are similar to ours. With other creatures, however, it is much more controversial. Could we build AIs that are conscious? I see no *a priori* reason why we couldn’t. Like many others, I do not see anything special about the way in which we are biologically constituted (cf. Kirk 2017, pp. 28–29; Graziano 2019). However, I am also sympathetic to the thesis that an AI program could not be conscious, as Christof Koch (2019) and John Searle (1980) hold, because programs would only be able to simulate consciousness. I agree with Koch that ‘experience does not arise out of computation’ (2019, p. xiv). A program that merely simulated the human brain, would not generate consciousness, because it does not generate the right causal structures of the brain—it just manipulates 0 s and 1 s.

I do not think much follows from this admission, however, because not all AIs are simply programs. Even AI robots of today exhibit complex causal structures. The difficult

question that faces us is ‘How can we tell whether an AI that possesses complex causal structures is conscious?’ I now turn to the question of animal consciousness, to show how the ‘Hard Problem’ of consciousness can be circumvented in that context, before turning to AI consciousness, where I argue that it cannot.

6 Nonhuman animal consciousness: circumventing the ‘hard problem’

It is common to think that nonhuman animals have conscious experiences like ours because of their similarity to us. Such a position is defended by Koch (2019, pp. 26–28), for example, who articulates three of the typical reasons that are given for attributing consciousness to nonhuman mammals. First, he claims, nonhuman mammals have a similar evolutionary history to us: we are similar, biologically speaking, to chimpanzees and other mammals. Second, the architecture of the nervous system is similar across mammals. Koch claims, ‘most of the close to nine hundred distinct annotated macroscopic structures that are found in the human brain are present in the mouse brain’ (2019, p. 27). And third, there are behavioural commonalities: nonhuman mammals grieve, express joy, and get angry just like we do.

Michael Tye (2017) employs similar reasoning in his defense of the claim that some nonhuman animals are conscious. Tye begins by noting that when we see another person apparently in pain (e.g., grimacing after falling over), we assume that the same causal story applies to that person as it would to us if we were in that situation. We thus take it for granted that such a person is in pain, as this is the simplest explanation of the common behaviour. Tye calls the principle behind this kind of approach ‘Newton’s Rule’, which he states as ‘The causes assigned to natural effects of the same kind must be, as far as possible, the same’ (2017, p. 72).

Tye thinks that Newton’s Rule gives us justification for thinking that other animals (including non-mammals) are conscious. Other animals’ (alleged) pain behaviours, for example, appear to be caused by the same kinds of processes that engender pain in us. Tye thinks that attributing consciousness to them is the simplest explanation; he claims ‘I am entitled to infer that the feeling of pain causes behaviour...in them too unless there is a *defeater*’ (2017, p. 75 emphasis in original). While Tye concedes that there are some significant differences between our brains and the brains of other creatures, he does not think these differences constitute defeaters. If a dog does not feel pain when it exhibits pain-like behaviour (e.g., after suffering a broken leg) we would need to explain why the dog lacks pain experience and a human would feel pain in the same case. We would need to posit some kind of special difference between humans and dogs. Given our close evolutionary history with

⁹ This is not to say that we are infallible, with respect to knowing what mental states we are in. See Andreotta (2019) for a discussion about the ways in which we go wrong in our mental state self-attributions.

dogs, and similar construction, this difference would appear quite mysterious, and so the claim that dogs are not conscious is not very parsimonious. Given how important sensory awareness is for *our* survival, it seems the same would be true of other species.

These considerations give us good grounds for attributing conscious experiences to nonhuman mammals. It is, however, one thing to make assumptions about the experiences that our mammalian cousins undergo, and another to make assumptions about animals that are quite different from us. Do fish, for example, feel pain in a similar way to us? Applying the kind of reasoning examined above, it may seem like they do. Consider a fish that has been placed on land after having been caught. Upon watching it gasp for air, as it flips about, it would be quite natural to think it is in pain. On the other hand, there are key differences between mammals and fish that give us reasons to be sceptical about such attributions. For example, it is commonly pointed out that fish do not possess a neocortex, as humans do. Given that the prefrontal cortex plays an important role in pain processing—that is, it is active during painful experiences (Ong et al. 2019)—this is thought to be a problem. Fish behaviour could simply be reflex-like, unaccompanied by any conscious experiences.

This is not a knock down objection, however. For one reason, as brains evolve, ‘newer’ areas may be capable of achieving the same function as ‘older’ areas (see Braithwaite 2010, pp. 12–13). Second, other creatures such as birds do not possess a neocortex, and yet many think that they contain the neurology capable of generating pain experiences. As Kenneth D. Harris points out, the ‘avian pallium [a part of the bird’s brain] contains circuits homologous to those of the mammalian neocortex’ (2015, p. 3185). And third, as Tye (2017, p. 81) points out, there is evidence that some humans who have been born without a human cortex can still undergo some experiences such as pain—that is, they are not completely in a vegetative state (cf. Shewmon et al. 1999).

In recent years, the view that consciousness is widespread has become widely accepted amongst experts. In 2015, at the Francis Crick Memorial Conference on Consciousness in human and nonhuman animals, the ‘Cambridge Declaration of Consciousness’ was devised by a prominent international group of cognitive neuroscientists, neuropharmacologists, neurophysiologists, neuroanatomists and computational neuroscientists. It says that:

The absence of a neocortex does not appear to preclude an organism from experiencing affective states. Convergent evidence indicates that nonhuman animals have the neuroanatomical, neurochemical, and neurophysiological substrates of conscious states along with the capacity to exhibit intentional behaviours.

Consequently, the weight of evidence indicates that humans are not unique in possessing the neurological substrates that generate consciousness. Nonhuman animals, including all mammals and birds, and many other creatures, including octopuses, also possess these neurological substrates

(Low et al. 2012, p. 2)

The declaration is noteworthy for two main reasons I wish to draw attention to. First, it recognises the scientific legitimacy of the thesis that consciousness is present in other creatures besides us. The second is the reasoning process underpinning this conclusion. We know that neurological substrates (e.g., the nervous system) generate consciousness in our own case. So, if there is evidence that other creatures possess similar neurological substrates, and behave in similar ways to us, then the best explanation of this observation is that they, too, have consciousness experiences. If an animal is different from us, however, we need to exercise caution. Michael Graziano, for example, exemplifies this approach in his scepticism of octopus consciousness. While not denying octopus consciousness, he says: that the ‘octopus nervous system is still so incompletely understood that we can’t yet compare its brain organization with ours and guess how similar it might be in its algorithms and self models’ (2019, p. 15).

As mentioned above, I think that this ‘human-comparative’ approach has yielded good results. Since there is nothing exceptional about the human brain (cf. Tononi and Koch 2015, p. 4), it is plausible to suppose other creatures with brains similar to ours will also be capable of undergoing experiences. Furthermore, this approach provides a methodology of determining whether a nonhuman creature is conscious. All we have to do is study the constitution and behaviour of an animal, and see how similar it is to our own. If it is similar, we should attribute consciousness to the creature. We do not need to solve the ‘Hard Problem’ of explaining why neurological substrates give rise to consciousness. Despite the success of this method, however, I will now suggest that it is limited when applied to AI consciousness.

7 AI consciousness

As difficult as it is to say how widespread consciousness is in the animal kingdom, determining whether a particular AI (or class of AIs) is conscious is a far more difficult task. One reason for this is that we will not be able to appeal to the same style of evolutionary argument that was advanced above, given that we are the ones creating the AIs. A sophisticated robot which behaves as if it is in pain, or expresses sorrow, may after all be a zombie that does not undergo any experiences. It may be that the robot was programmed to

imitate human emotions to aid human–robot interactions. Since the AIs of today are already capable of emulating basic human-like speech or behaviours, it is conceivable that AIs could one day perfectly mimic our behaviour without undergoing any experiences. So, the question is ‘How can we tell the difference between a zombie AI and a conscious AI?’ In the subsections that follow, I critique several approaches.

7.1 The argument from functionality

Applying Newton’s rule (see Sect. 6), Tye thinks we should attribute consciousness to an AI if the AI’s behaviour resembles our own, and if the causal mechanisms that give rise to such behaviour operate or function in a way that resembles our own when we undergo similar experiences. For example, if we can determine that an AI’s pain behaviours are ‘caused by tissue damage’ (2017, p. 193), or if an AI’s ‘pain diminishes the desire to eat’ (2017, p. 193), then we have evidence in Tye’s view that the AI functions as we do. It does not matter to Tye whether an organism is part of our evolutionary lineage or whether it is ‘artificial’ (2017, p. 190). What matters is how it functions.

To illustrate this idea, Tye considers the case of Commander Data from *Star Trek*, who is described as a fully functioning android. (See also Dahj and Soji—Data’s android descendants featured in the recent TV series *Star Trek: Picard*.) Tye stipulates that Data behaves just like a human would: he exhibits human-like pain behaviour when he suffers damage to his body, and can verbally express human-like emotions. Tye notes that while Data’s brain is different to ours—it is a positronic brain formed from an alloy of platinum and iridium—it is *functionally* equivalent to ours, making Data a functional isomorph of a human. Tye concludes that there is ‘evidence that he feels anger, fear, and pain’ (2017, p. 179). A reconstruction of Tye’s reasoning about Data, which incorporates Newton’s Rule, can be offered as follows:

[P1] AIs that function and behave in the same way that humans do are most likely conscious.

[P2] Commander Data functions and behaves in the same way as humans do.

[C] Commander Data is most likely conscious.

I see two main problems with this argument. Firstly, [P1] assumes that there is nothing significant about the material underlying the neurological substrates that give rise to consciousness in us, or in closely related creatures (e.g., dogs, dolphins). According to [P1], it does not matter what an AI is made of—all that matters is how the AI functions. I am of course not claiming that consciousness can only arise from biological creatures like us; what I am claiming is the only evidence we have of consciousness arising is from creatures

like us. The worry I have with [P1] is a justificatory one. If we start from the observation that the neurological substrates we possess give rise to consciousness, and then observe that other animals share similar neurological substrates; we may fairly infer that those creatures are conscious. But what justification or evidence can be brought to bear upon the claim that all creatures that function, and behave, like we do—regardless of material constitution—are conscious?

7.2 The silicon chip argument

One way to justify the thesis that AIs that are constituted differently to us, yet function as we do, are conscious is the *silicon chip argument* (see Searle 1992; Chalmers 1996; Tye 2017; Schneider 2019). The argument involves a thought experiment which requires us to imagine that a single neuron in a person’s brain, which we can refer to as *A*, is replaced by a silicon chip which we can refer to as *A**. It is stipulated that *A** performs the exact same local function as *A*, and is connected to the same neurons that *A* was. Furthermore, the input/output function of *A* is equivalent to *A**.

The first step of the argument is to consider what the end result of this procedure would feel like from the first-person perspective. Intuitively it would seem like nothing would change: replacing *A* with *A** would have no noticeable effect—given that there are around 100 billion neurons in the brain. The next step in the argument is to imagine what would happen if a second neuron, *B*, was replaced by a silicon chip *B**. Would there be a change? Again, the intuition here is that there would still be no change. We are then to imagine that more and more neurons are swapped until finally all are replaced. These duplicate neurons would have, as Susan Schneider says, ‘every causal property of neurons that make a difference to your mental life’ (2019, p. 28).

Like many, I find it plausible that consciousness would still be intact after the procedure. However, this is not the only explanation I find plausible. Another is that the subject may still have experiences, but they would differ from what they normally experience (Chalmers 1996, p. 263). Instead of lemons tasting sour, for example, they could taste like vinegar. A third possibility is that one’s experiences will fade, as neurons are replaced, until eventually all consciousness is lost (Chalmers 1996, p. 251). These latter possibilities are dismissed by Chalmers (1996), Schneider (2019), and Tye (2017) because of their purported implausibility. They hold the view that there would be no felt change.

In response to these authors, I would say that the two alternative possibilities only seem implausible because of the implicit assumption that is underpinning their intuitions—namely, that all causal sequences that mirror the ones underlying our conscious experiences will also result in consciousness. In other words, the constitution of the materials involved in these causal sequences do not matter. If this

assumption is granted, then of course there is every reason to favour the interpretation which says that silicon replacement surgery would not result in the loss of consciousness. What reasons do we have to justify the assumption though? The above authors appear to rely on intuition, and the idea that the other alternatives seem too implausible. However, I think that intuition may provide limited guidance here, given that it is still not understood why certain brain states gives rise to certain experiences.

In principle, the assumption in question could be empirically tested—without a solution to the ‘Hard Problem.’ One way would be to assess the level of consciousness at each step of the procedure. A subject could be asked how each small change made a difference to their consciousness. This method would not be flawless, however, because there remains the possibility that as neurons are replaced, consciousness is lost, while the behaviour and communication abilities remain (see Searle 1992, pp. 66–67). There are, however, practical barriers that prevent such a test from being administered any time soon: first, because it is far beyond our technological capabilities, and second because of ethical concerns. We must rely on our intuitions here then, for the time being, given that there remains much that is mysterious about the way that consciousness arises from the brain.

7.3 Epistemological problems with identifying functional analogues

Let us assume for the sake of argument that [P1] is true: Suppose that AIs that function and behave in the same way that humans do are conscious. Even if this claim is true, however, we are still left with the practical problem of knowing when an AI’s functionality and behaviour is similar enough to ours to warrant attributing consciousness to it. Even if we grant that Commander Data functions and behaves like we do, and is thus conscious, how would we know whether an AI that we have built is conscious if it functions very differently to us? Tye is, after all, careful to stipulate that Data is virtually identical to a human, in terms of his behaviour and functionality. This stipulation allows Tye to make a similar kind of inference that was used above with respect to animal consciousness—namely, if it functions and behaves like us, it is likely conscious. The problem that faces us, however, is that functional isomorphs like Data, or silicon chip replacement surgeries, are not likely to occur any time soon. (Though humanlike automata are widely ‘discussed and pursued’ as Miller (2017, p. 3) says). The AIs that we are currently building, and will continue to build, are quite different from us in terms of their function and constitution.

Consider, for example, H-1—an autonomous humanoid robot developed by Gordon Cheng and his team at Technical University of Munich (Cheng et al. 2019). While similar

to other humanoid robots, e.g., Boston Dynamic’s Atlas or Honda’s ASIMO, H-1 is unique in that it is equipped with artificial robot skin. This skin consists of over 1200 cells, and has over 13,000 sensors over its body (Technical University of Munich (TUM) 2019). This artificial skin enables H-1 to ‘feel’ and ‘respond’ to the world, by detecting pressure on its body and the temperature around it. Like the species that exist on Earth, this sensory capacity enables H-1 to protect itself from harm; and it can also help it not to accidentally hurt the humans it interacts with.

As impressive as H-1 is, its number of sensors do not compare to the 5 million receptors human skin has (Technical University of Munich (TUM) 2019). In addition, while H-1 resembles us in several interesting ways, H-1 is too different to us, in terms of behaviour and functionality, for anyone to claim that it consciously feels (e.g., pain) like we do. Why is this? For one thing, the way that H-1 ‘learns’ about its environment is different from the way that we do: it does not possess a brain anything as complex as the 100 billion neuron brain that we have—the source of our conscious experiences; neither does H-1 exhibit behaviour like ours when it receives damage to its body. Furthermore, H-1 does not possess the capacity for language. So, applying a method such as Newton’s rule, it is implausible to say that it is conscious.

Might H-1 eventually possess the capacity for consciousness if future incarnations of it become more complex? This is a hard question. I argued above that complexity alone is not sufficient for conscious experience. So even if H-1 was equipped with 5 million skin receptors, I do not think this would guarantee it. A more complex version of H-1 may, however, have the capacity for consciousness. The problem that we will face is how will we recognise it, if its functionality and behaviour do not resemble our own. It would be humancentric to suppose that only close functional isomorphs can be conscious. There could be multiple ways in which consciousness could arise in a system. Neurons interacting in the ways that they do in the human brain may be one way, but there could be countless others.

7.4 Schneider’s AI consciousness test

One way to circumvent these problems is to come up with a test—one that would allow us to bypass the ‘Hard Problem’ of consciousness. Susan Schneider (2019) has recently proposed one, which she calls the AI Consciousness Test (ACT). The test is designed to reveal whether an AI is conscious—regardless of its constitution, or how it functions. The ACT can, according to Schneider, distinguish between ‘a creature that merely has cognitive abilities, yet is a zombie’ (2019, p. 51) and a genuinely conscious creature. The ACT is linguistic based, like Turing’s famous test: it requires a human to ask an AI a series of

questions. Unlike Turing's test, however, it attempts to identify the presence of a conscious mind. Schneider's idea is that these questions would not be answerable by an AI zombie. Some of the examples from Schneider's list of sample questions include: 'Could you survive the permanent deletion of your program? What if you learned this would occur?... What is it to be like you right now?' (2019, p. 55). In Schneider's view, a 'satisfactory response to one or more of the...questions or scenarios is sufficient for passing the test' (2019, p. 54).

When applied to current technology, the ACT seems to produce the right result. The iPhone assistant Siri, for example, when asked 'What is it to be like you right now?' responds by saying 'I'm not sure I understand'. This is a good result, as it is implausible that current smart phone assistants are conscious. The same is true with respect to other AIs (e.g., robot vacuum cleaners and self-driving cars) that exist today. Despite such results, the ACT does not, to my mind, constitute a good test. There are two objections to it that I provide below.

The first is that the test does not seem sufficient for establishing that a creature is conscious. Suppose that we invent superintelligent AIs, and suppose further that we program them to care about their own survival. One unintended consequence—or a 'perverse instantiation' as Bostrom puts it (2014, p. 146)—of this goal might be that deception is sometimes required, if it increases the chances of survival. A superintelligent AI zombie could learn from books, academic articles and blogs that humans would care about it in a greater way if they believed it was conscious. Given that adult humans could easily pass the ACT, the AI could simply learn the right responses and avow them when asked. Schneider (2019, p. 53) thinks that we can get around the problem by boxing the AI in—that is, not allowing it to get information from the outside world, such as from books or the internet. Schneider thinks that if the ACT was administered at the R & D stage, the AI would not have access to such information and be unable to fake the test.

One immediate problem with Schneider's idea is that it places a large amount of pressure on the creators of AIs to administer the test at the right time. For the reasons just given, the test becomes unreliable after the AI is 'released', or exposed to information. As soon as the AI is able to access information from the world, it become hard to know if the AI is 'faking' it. Still, even if the ACT could be administered during the R & D stage, a zombie AI may pass the test. Since the test requires a sophisticated AI (it needs to be able to communicate with us), it may be that part of its learning, in the R & D stage, involved the study of human language. In the process, the AI may have learned some facts about humans, such as what motivates them. Given our current concerns about deep learning—e.g., the black box problem—it may not even be clear to the designers of

the AI how such learning was achieved. These concerns cast doubt upon the ACT's ability to identify a conscious entity.

The second objection that I will raise is that the ACT is too limited in its application. Even if we agree that passing the ACT is enough to establish the presence of a conscious being, it is easy to imagine cases, where the test fails to identify a conscious AI—namely, in cases where the AI lacks the capacity for language. Schneider is aware of this problem and concedes that the test is only a sufficient test. However, this concession raises a bigger problem not only for her test, but for linguistic tests in general—namely, that such tests will fail to identify consciousness in creatures without language, such as animals, or small children.

Furthermore, linguistic tests will also fail in cases, where a creature has a mind, and could potentially learn language, but lacks the ability to express language. Consider the phenomenon of locked-in syndrome in humans. These are people who are conscious but have no means of speaking or communicating. The possibility of this occurring in humans raises the possibility of us inadvertently creating conscious AIs who can feel pain but do not possess a capacity for communication or movement. In the anticipation of such objections, Schneider claims that her test is an 'initial step' (2019, p. 49). My objection has been that the ACT sets the bar too high, and thus will only be able to identify a restricted subset of conscious AIs that are deserving of moral concern.

7.5 Integrated information

I have argued that a human-comparative approach to identifying consciousness in other creatures is limited in several ways. The approach may work well when the creature in question is similar to us; but, I have argued, the approach will be stretched when it comes to AIs that are different to us, yet are capable of experiences (e.g., the ability to suffer). One promising theory of consciousness that is relevant to this problem is the Integrated Information Theory (IIT)—most notably defended by Giulio Tononi (2004, 2012). According to this theory, the amount of integrated information present in an object or system—whether this be a human brain, dolphin brain or machine—depends on whether that object or system is conscious. Tononi measures the amount of integrated information that a system is capable of in terms of its Φ value. As the Φ value increases, the theory holds, the amount of consciousness increases. Tononi claims that ' Φ is the amount of causally effective information that can be integrated across the informational weakest link of a subset of elements' (2004, p. 1).

One benefit of IIT is that it is empirically testable. The IIT says that the loss of consciousness should be correlated with a low Φ value—such as during sleep or comatose states—and should be higher when consciousness returns. This turns out to be the case (Tononi and Koch 2015, p. 9). IIT should

also be able to identify the presence of consciousness in subjects who are suffering from locked-in syndrome—subjects who cannot move or communicate but are aware of their surroundings. If the subject is still conscious, the Φ value should still be high and thus able to be investigated by science. Such characteristics of the theory make it advantageous over some of the other tests for consciousness we have examined. For one reason, human-like behaviour or communication is not required to identify consciousness: all that is required is the presence of integrated information. And neither does the ‘Hard Problem’ have to be solved. We do not need to know *why* integrated information gives rise to consciousness.

While the IIT is an important theory, I do not think it will fully resolve the question of AI consciousness if AIs are built that are different to us. One reason is that, like the other theories examined, the theory’s starting point is to match human experience with certain correlates: in this case integrated information. This makes the theory ideal for testing in our own case. However, the theory is not as easily testable with non-human creatures. After all, communication is required to verify that consciousness really is correlated with a high Φ value. There remains the possibility that a high Φ value may be sufficient but not necessary for consciousness. For instance, how would we test whether an AI that registered a low Φ really was unconscious, after it claimed to be, and vice versa (a high Φ value but no external signs of consciousness)? A proponent of IIT who is convinced of the truth of theory may claim that if a low Φ value in humans indicates a lack of conscious, then a low Φ value in any system indicates a lack of consciousness. However, it remains possible that there are other ways of achieving consciousness that do not involve integrated information.

My point here is not to undermine IIT by finding logically possible problems with it. My point is, rather, that some of the advantages of the IIT—such as its ability to be corroborated with human experience—means that as we step away from humans our capacity to corroborate diminishes. When we step away from our own consciousness and attempt to detect its presence in other creatures, the same kinds of problems that were identified in the other accounts examined arise for the IIT.

8 Conclusion

Progress can and has been made on the problem of animal consciousness, and in turn animal rights, without the possession of a solution to the ‘Hard Problem’ of consciousness. By comparing animals’ behaviours, and the internal mechanisms that give rise to those behaviours, with our own we can make well-grounded assumptions about what their mental lives are like. In terms of rights, it does not matter

how intelligent a creature is, or how much empathy we feel for that creature. What matters is whether they can experience pleasure or pain.

While this ‘human-comparative approach’ has clearly produced important results, I have argued that it is of limited use when applied to the problem of AI consciousness, and in turn AI rights. For one reason, we will not be able to appeal to a shared evolutionary history, given that we are the ones creating AIs.¹⁰ For a second reason, the AIs that we are currently building, and will most likely continue to build, are constituted from material which is different to us. And lastly, AIs are likely to function, at least for the foreseeable future, in ways that differ greatly from to us.

A solution to the ‘Hard Problem’ would of course make this problem much more tractable. If we knew why certain brain states are accompanied by experience, or why certain brain states give rise to a color experience versus a pain experience, then we would be better placed to answer the question of why certain configurations of matter give rise to experience or why others do not. As it stands, we do not have a complete understanding of why brain states are accompanied by certain experiences, and that makes progress on the question of AI rights difficult. One upshot of this discussion is that I do not think that the ‘Hard Problem’ should be dismissed as easily as it has been by some, such as Caruthers, who claims ‘The “hard problem” of consciousness in *humans* has been overblown’ as (2019, p. x emphasis in original). I have tried to show why the problem of AI rights makes this interpretation of the ‘Hard Problem’ unjustified. I contend that we should be focused on trying to solve it for the practical reasons laid out here.

In claiming that consciousness should ground AI rights, I am not as Gunkel claims ‘closing down critical inquiry because such questions can be easily dismissed as futuristic and not very realistic’ (2018, p. 95). On the contrary, there is much to inquire about the nature of consciousness, and such inquiries can be undertaken now. Since I have argued that we cannot begin to address the question of a creature’s rights until we have some sort of understanding of whether that creature has experiences, it becomes a question that is critical to solve for practical reasons. After all, if we do not understand how consciousness arises, we may inadvertently create creatures that have it, and treat them in ways that would generate great amounts of suffering without ever knowing we are doing so. After all, we already inflict large amounts of pain and suffering on moral agents today (Torrance 2011, p. 133).

¹⁰ While we may not be passing on our own genetic information to AIs, we can still pass on (sometimes unwillingly) elements of ourselves. Consider, for example, the problem that certain biases—e.g., racist, sexist, or ageist—have crept into AI algorithms (O’Neil 2016).

An implication of my main conclusion is that the creators, investors, engineers and governments who are seeking to build complex AIs, as well as society at large, should not limit their concerns to ones involving our own wellbeing. If we create conscious AIs (whether intentionally or inadvertently) we will need to take into account their interests, e.g., by attempting to avoid inflicting unnecessary suffering on them—the same goal that is sought by proponents of animal rights. The question of what resources should be allocated to achieve such ends is a complicated one, and will depend on what kinds of AIs we create. The decision to make AIs that are conscious is not one that should be taken lightly.

If we do decide to make conscious AIs, and cannot solve the ‘Hard Problem’ of consciousness, then our best chance of knowing that we have done so, I have argued, is to attempt to make them as close to us as possible. By copying, and modifying, designs of conscious creatures that already exist in nature (such as humans, or closely related species) it may be possible to create novel species who exhibit intelligent behaviours, perhaps even superior to our own. In the same way that the behaviours of domesticated animals differ from their wild cousins (compare the differences in aggression between wild cats and domesticated cats), it may be possible to genetically modify, or select for intelligence. This route to AI consciousness is not without its problems, however. Not only is such an approach beyond our current technological reach, but there are many ethical issues with such kinds of research, and also motivational questions about why we would want to undertake such a project in the first place. There is also the further problem, or paradox, that the very act of creating conscious AIs (including the research and development phases) could not be carried out with the informed consent of such AIs, as informed consent first requires consciousness (Miller 2017). Given that informed consent is taken very seriously with respect to human agents in medical contexts, this is a significant concern.

If we wish to avoid such problems, then it may be best to avoid creating conscious AIs altogether by pursuing AI designs that are as different from us as possible. For the reasons given in this paper, this will not guarantee that we have avoided creating conscious AIs, and thus have done our best to reduce unnecessary suffering; but it may be the best that we can hope to achieve, as long as the ‘Hard Problem’ of consciousness remains unsolved.

Acknowledgements Earlier versions of this paper were presented to audiences at the annual Limina Conference, HUMANIFESTO, at the University of Western Australia in July 2019; and at Curtin University, Bentley, in early 2020. I thank everyone who took part in those enlightening discussions. I am also grateful to Sean Ramsey, who read several draft versions of the paper and offered helpful suggestions and criticisms. I also would like to thank two anonymous referees of this journal for their insightful comments.

Funding Not applicable.

Compliance with ethical standards

Conflicts of interest All authors declare that they have no conflict of interest.

References

- Andreotta AJ (2019) Confabulation does not undermine introspection for propositional attitudes. *Synthese*. <https://doi.org/10.1007/s11229-019-02373-9>
- Baron-Cohen S (2011) *The science of evil: on empathy and the origins of cruelty*. Basic Books, New York
- Batson CD, Klein TR, Highberger L, Shaw LL (1995) Immorality from empathy-induced altruism: when compassion and justice conflict. *J Pers Soc Psychol* 68:1042–1054. <https://doi.org/10.1037/0022-3514.68.6.1042>
- Bentham J ([1789] 2005) *An introduction to the principles of morals and legislation*. In: Burns JH, Hart HL (eds). Oxford University Press, Oxford
- Bloom P (2016) *Against empathy: the case for rational compassion*. Ecco, New York
- Bostrom N (2014) *Superintelligence: paths, dangers, strategies*. Oxford University Press, Oxford
- Braithwaite V (2010) *Do fish feel pain?*. Oxford University Press, Oxford
- Bryson JJ (2010) Robots should be slaves. In: Wilks Yorick (ed) *Close engagements with artificial companions: key social, psychological, ethical and design issues*. John Benjamins Publishing Company, Philadelphia, pp 63–74
- Burdon PD (2012) A theory of earth jurisprudence. *Australian J Legal Philos* 37:28–60
- Carruthers P (1992) *The animals issue*. Cambridge University Press, Cambridge
- Carruthers P (2019) *Human and animal minds: the consciousness questions laid to rest*. Oxford University Press, Oxford
- Chalmers DJ (1996) *The conscious mind. In search of a fundamental theory*. Oxford University Press, Oxford
- Chalmers DJ (2010) *The character of consciousness*. Oxford University Press, Oxford
- Chapron G, Epstein Y, López-Bao JV (2019) A rights revolution for nature. *Science* 363(6434):1392–1393. <https://doi.org/10.1126/science.aav5601>
- Cheng G, Dean-Leon E, Bergner F, Rogelio J, Olvera G, Leboutet Q, Mittendorfer P (2019) A comprehensive realization of robot skin: sensors, sensing, control, and applications. *Proc IEEE* 107(10):2034. <https://doi.org/10.1109/JPROC.2019.2933348>
- Coeckelbergh M (2012) *Growing moral relations: critique of moral status ascription*. Palgrave MacMillan, New York
- Darling K (2017) “Who’s Johnny?” Anthropomorphic framing in human-robot interaction, integration, and policy. In: Lin P, Jenkins R, Abney K (eds) *ROBOT ETHICS 2.0: from autonomous cars to artificial intelligence*. Oxford University Press, Oxford, pp 173–191
- Dennett DC (1996) *Kinds of minds: toward an understanding of consciousness*. Basic Books, New York
- Dennett DC (2018) Facing up to the hard question of consciousness. *Philos Trans R Soc B: Biol Sci*. <https://doi.org/10.1098/rstb.2017.0342>
- Frankish K (2016) Illusionism as a theory of consciousness. *J Conscious Stud* 23:11–39

- Garner R (2013) A theory of justice for animals: animal rights in a nonideal world. Oxford University Press, New York
- Godfrey-Smith P (2016) Other minds: the octopus, the sea, and the deep origins of consciousness. Oxford University Press, Oxford
- Goertzel B (2002) Thoughts on AI morality. In: Dynamical psychology: an international, interdisciplinary journal of complex mental processes (May). <https://goertzel.org/dynapsyc/2002/AIMorality.htm>
- Graziano MSA (2019) Rethinking consciousness: a scientific theory of subjective experience. W. W. Norton, New York
- Gunkel DJ (2012) The machine question: critical perspectives on AI, robots and ethics. MIT Press, Cambridge
- Gunkel DJ (2014) A vindication of the rights of machines. *Philos Technol* 27:113–132
- Gunkel DJ (2018) Robot rights. MIT Press, Cambridge
- Harris KD (2015) Cortical computation in mammals and birds. *Proc Natl Acad Sci* 112:3184–3185. <https://doi.org/10.1073/pnas.1502209112>
- Herzog H (2010) We love, some we hate, some we eat: why it's so hard to think straight about animals. Harper Collins, New York
- Hohfeld W (1919) Fundamental legal conceptions as applied in judicial reasoning. Yale University Press, New Haven
- Howe A (2019) Why kant animals have rights? *J Anim Ethics* 9:137–142
- Kant I (1963) Lectures on ethics, trans. Infield L. Harper Torchbooks, New York
- Kirk R (2017) Robots, zombies and us: understanding consciousness. Bloomsbury Publishing, New York
- Koch C (2019) The feeling of life itself: why consciousness is widespread but can't be computed. MIT Press, Cambridge, MA
- Kramer MH (1998) Rights Without Trimmings. In: Kramer M, Simmonds N, Steiner H (eds) A debate over rights. Oxford University Press, Oxford, pp 7–111
- Low P, Panksepp J, Reiss D, Edelman D, Van Swinderen B, Koch C (2012) The Cambridge declaration on consciousness. <http://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf>. Accessed 24 Mar 2020
- Miller LE (2017) Responsible research for the construction of maximally humanlike automata: the paradox of unattainable informed consent. *Ethics Inf Technol*. <https://doi.org/10.1007/s10676-017-9427-3>
- O'Donnell E, Talbot-Jones J (2017) Three rivers are now legally people—but that's just the start of looking after them. *Conversation*. <https://theconversation.com/three-rivers-are-now-legally-people-but-thats-just-the-start-of-looking-after-them-74983>. Accessed 24 Mar 2020
- O'Neil C (2016) Weapons of math destruction: How big data increases inequality and threatens democracy. Crown, New York
- Ong W-Y, Stohler CS, Herr DR (2019) Role of the prefrontal cortex in pain processing. *Mol Neurobiol* 56:1137–1166. <https://doi.org/10.1007/s12035-018-1130-9>
- Prinz L (2011) Against empathy. *South J Philos* 49(Spindel Supplement):214–233
- Prinz J (2012) The conscious brain: how attention engenders experience. OUP USA, New York
- Putnam H (1964) Robots: machines or artificially created life? *J Philos* 61:386–407
- Rawls J ([1971] 1999) A theory of justice (Revised Edition). Harvard University Press, Harvard
- Rühs N, Jones A (2016) The implementation of earth jurisprudence through substantive constitutional rights of nature. *Sustain: Sci Pract Policy* 8(2):174. <https://doi.org/10.3390/su8020174>
- Russell S (2019) Human compatible: artificial intelligence and the problem of control. Penguin Books, London
- Schneider S (2019) Artificial you: AI and the FUTURE OF YOUR MIND. Princeton University Press, Princeton
- Schwitzgebel E, Garza M (2015) A defense of the rights of artificial intelligences. *Midw Stud Philos* 39:89–119
- Searle JR (1980) Minds, brains, and programs. *Behav Brain Sci* 3:417–457
- Searle JR (1992) The rediscovery of the mind. MIT Press, Cambridge
- Shewmon A, Holmes G, Byrne P (1999) Consciousness in congenitally decorticate children: developmental vegetative state as self-fulfilling prophecy. *Dev Med Child Neurol* 41(6):364–374
- Singer P ([1975] 2009) Animal liberation: the definitive classic of the animal movement (Updated Edition). Harper Collins Publishers Inc, New York
- Slote M (2010) Moral sentimentalism. Oxford University Press, Oxford
- Technical University of Munich (TUM) (2019) Biologically-inspired skin improves robots' sensory abilities. *Science Daily*. www.sciencedaily.com/releases/2019/10/191010125623.htm. Accessed 9 Jan 2020
- Tegmark M (2017) Life 3.0: being human in the age of artificial intelligence. Alfred A. Knopf, New York
- Tononi G (2004) An information integration theory of consciousness. *BMC Neurosci* 5:42. <https://doi.org/10.1186/1471-2202-5-42>
- Tononi G (2012) Phi: a voyage from the brain to the soul. Pantheon Books, New York
- Tononi G, Koch C (2015) Consciousness: here, there, and everywhere? *Philos Trans R Soc B* 370:20140167. <https://doi.org/10.1098/rstb.2014.0167>
- Torrance S (2011) Machine ethics and the idea of a more-than-human moral world. In: Anderson M, Anderson SL (eds) Machine ethics. Cambridge University Press, Cambridge, pp 115–137
- Turing AM (1950) Computing machinery and intelligence. *Mind* 49:433–460
- Tye M (2017) Tense bees and shell-shocked crabs: are animals conscious?. Oxford University Press, Oxford
- Wallach W, Allen C (2009) Moral machines: teaching robots right from wrong. Oxford University Press, Oxford
- Wenar L (2005) The nature of rights. *Philos Public Aff* 33:223–252. <https://doi.org/10.1111/j.1088-4963.2005.00032.x>
- Xu X, Zuo X, Wang X, Han S (2009) Do you feel my pain? Racial group membership modulates empathic neural responses. *J Neurosci* 29:8525–8529

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.