Suffer the robot? Pathways to artificial moral patiency (8229 words)

Dr Henry Shevlin, Leverhulme Centre for the Future of Intelligence

Introduction

Even a casual viewer or reader of science-fiction will doubtless have encountered the image of the maltreated machine. From the abused simulants of *Blade Runner* to the neglected child-robot 'David' in *A.I. Artificial Intelligence*, we seem to have little difficulty in imagining artificial beings feeling pain and distress, and even empathising with them. The kind of advanced artificial intelligence we see in fiction is of course far removed from the capabilities of real world machines. However, in light of the breakneck pace of recent developments in machine learning, it does not seem wholly fanciful to speculate that within a matter of decades we may be able to build artificial beings with cognitive capacities approaching those of animals and humans. It is natural to wonder, then, whether such machines – like their counterparts in fiction – might one day come to possess some moral status.

This paper will explore this possibility, and consider whether and how we might one day have grounds for classifying artificial beings as having moral status, specifically what I term moral patiency. I begin in Section 1 by spelling out the notion of a moral patient, distinguishing it from the notions of moral relevance and moral obligation. In Sections 2-4, I examine three pathways by which we might rationally decide to extend moral consideration to artificial beings, namely suffering, preferences, and personhood, noting the challenges they face for their practical employment as tests of machine moral patiency. In Section 5, I lay out an alternative proposal for establishing moral patiency in artificial systems called the *Biological Analogy* approach, which suggests at least some questions of machine patiency can be answered by reference to existing frameworks for determining animal moral patiency. I go on in Section 6 to give brief summary of some of the limitations of existing artificial systems, particularly in regard to general intelligence, and argue that by the lights of the Biological Analogy approach, none yet come even close to meeting criteria for moral patiency. I conclude in Section 7 by noting that while machine moral patients may be a long way off, the questions of how they might come about and how we might recognise them when they do are crucial and fruitful lines of enquiry.

1. Moral relevance and moral patiency

To begin with, it will be helpful to explain precisely what I have in mind with the notion of moral patiency.[1] Roughly speaking, I take the notion of a moral patient to that of a being that can be benefitted or harmed in a sense potentially relevant for moral decision-making. More specifically, I will count as a moral patient any being with *intrinsic interests*; that is, interests that it possesses not in virtue of its relationship to a specific moral agent or a moral community, but rather in virtue of its internal cognitive or biological features. Such harms may might take the form of emotional or physical suffering, but depending on one's broader ethical theory, might also involve non-experiential harms, such as being deprived of one's

[1] The questions of how to analyse notions such as moral patiency and moral status have been discussed widely by philosophers. My treatment of it here context is deliberately brief and consequently somewhat cursory. For more detailed discussion, see, e.g., Korsgaard (1996), Kamm (2007), and McMahan (2002).

privacy, autonomy, or the fruit of one's labour. I take it that beings with moral patiency in this sense include, at the very least, all sentient humans, but most would wish to include many non-human animals as moral patients.

This account of a moral patient as a being that can be harmed in some (yet to be specified) relevant moral sense is deliberately broad and vague; I am, at least at the outset, casting a deliberately wide net so as to accommodate as wide a range of different normative ethical approaches as possible. However, I do wish to distinguish moral patiency from two related notions. The first is that of *intrinsic moral relevance*. While I take moral patients to be examples of beings with intrinsic moral relevance, there may be things that possess it without qualifying as moral patients. I have in mind, for example, entities such as great works of art, non-sentient natural objects like rivers and mouintains, and sacred artefacts. Which such things undoubtedly have instrumental value, insofar as they provide pleasure and others benefits to sentients being, one might also think that they are 'final goods' (Korsgaard, 1983) and matter in their own right. They do not qualify as moral patients in the way that I using the term, however, insofar as they cannot properly be said to be individually benefitted or harmed.[2]

A further distinction worth drawing is the one between the notions of moral patiency and moral obligation. While some might endorse the idea that we hold moral obligations towards beings just in virtue of the qualities that make them moral patients (Rachels, 1991), this is not the only way one might conceive of how the two ideas are related. Many think, for example, that our special relationships to certain moral patients – our family, our friends, our pets, and perhaps our species – give us enhanced moral obligations towards them (see, e.g., Scanlon, 1998, and Steinbock, 2011). One might also think it possible to recognise someone as a moral patient (that is, as an entity able to be harmed in a potentially morally relevant sense) while rejecting the idea that one actually has any obligations to them. To give an extreme example, a moral community might believe it has no obligations towards outsiders while nonetheless recognising them as fully capable of suffering and being otherwise harmed. Likewise, a retributive judge might determine that a heinous criminal has waived all right to moral treatment by others. There is thus room for slippage between the notions of being a moral patient and being an entity to whom we bear moral obligations.

With these distinctions in place, I can clearly state the question to be addressed in what follows, namely under what circumstances an artificial system could reasonably be recognised as an entity capable of being benefitted and harmed in a morally relevant sense. This is distinct from the question of whether an artificial system could come to have moral relevance in a broader sense (as might be the case for a particularly beautiful or significant technological artefact), as well as the question of whether we could have moral obligations towards machines.

A final point to note is that I will say nothing in what follows about *degrees* of moral patiency. It seems reasonable to most of us, for example, to think that while humans and mice are both moral patients, the kind of harms that can be inflicted to a human are much more ethically serious than those that can inflicted upon a mouse, giving rise to differences in

---

[2] The distinction I draw here between morally relevant things that are moral patients and those that are not corresponds roughly to the distinction drawn by Kamm between things of intrinsic value *for whose sake* we can act and those for whom we cannot. As he puts it, "I do not act for its sake when I save a work of art, because I do not think of its good and how continuing existence would be good for it when I save it" (Kamm, 2007: 228).

moral obligation. While this question of degree of moral patiency is a critical one for the broader project of providing a full account of the kinds of obligations we might owe to artificial beings, I will set it aside for the purpose of focusing on the specific issue of how we might recognise an artificial being as having (any) status as moral patient.

2. Pathways to moral patiency (1): suffering

Let us consider, then, what intrinsic properties or capacities we might be able to identify that would ground an artificial system's status as a moral patient.[3] One promising route in this regard come from the notion of suffering. Even if we do not wish to claim that this is all there is to moral patiency, certainly it is natural to think that if a being can suffer, it has at least some interests, viz. not suffering, and thus qualifies as a moral patient in the sense described above. Applying this to artificial entities, then, the notion of suffering might provide a conceptually straightforward pathway to moral patiency: if and when we build machines capable of suffering, we will have built moral patients.

There are several attractive features of this approach. First, as noted, the claim that suffering grounds at least some minimal form of moral patiency is relatively uncontroversial; even if we wish to allow there is more to well-being than pain and distress, it is hard to deny that such forms of suffering are often detrimental to our interests and those of non-human animals. Second, a suffering-based approach to moral patiency can accommodate quite simple forms of sentient being. Thus while we may not wish to consider fish or lobsters *persons* or as having interests connected to autonomy or personal fulfilment, if it could be established that they suffer this would give us some grounds for potentially taking their interests into account. This latter point is of special relevance to machine patiency insofar as it seems antecedently plausible that long before we have achieved human-level artificial intelligence with claims to more robust moral notions such as personhood or self-determination, we will along the way build systems with the cognitive complexity equivalent to that of non-human animals that might potentially qualify as moral patients in virtue of capacities to experience pain, fear, or other simple forms of suffering. Finally, a suffering-based approach to moral patiency be at least in principle empirically tractable, insofar as it identifies moral patiency with the capacity to undergo specific kinds of psychological state which we might expect to be amenable to assessment via the tools of cognitive science.

Despite this superficial simplicity, however, suffering-based approaches to moral patiency face serious challenges of both a philosophical and practical character. To give an example of the former, consider the difficulty involved in determining which specific states qualify as instances of suffering. On the one hand, we can all identify certain canonical cases of suffering, such as intense pain, nausea, and hunger, or extremes of negative emotions like grief and anxiety. There are many cases at the margins, however, that are harder to gauge. Should we count the pain of a stubbed toe, the mild annoyance of missing breakfast, or the boredom of a long bus trip instances of suffering, or are they too weak and ephemeral to qualify?

Even if we can agree on broad conceptual criteria for suffering, however, the practical

---

[3] As with the discussion of the concept of moral patiency, the question of possible *grounds* of moral patiency has been met with many different responses (see Jaworska & Tannenbaum, 2018). The three possibilities considered here are thus far from exhaustive.

assessment of whether a non-human entity is undergoing suffering in a given case is highly challenging even in more familiar domain of animal welfare research (see, e.g., Sneddon, Elwood, Adamo, & Leach, 2014). We can think of a capacity for suffering as involving at least two conditions. First, for a being to suffer it must be capable of undergoing at least some *negatively valenced psychological states*. Examples of such negatively valenced states in humans include pain, hunger, nausea, and a variety of forms of emotional distress. Second, in order to qualify as suffering, these states must also be consciously experienced by the being in question. This may seem an unnecessary point, given that states such as pains are frequently assumed to be conscious by definition, but note that we cannot rule out the possibility that (especially in non-human cases) some negatively valenced states might be experienced unconsciously, and hence not lead to suffering in the sense under discussion.

Consider first the problem of determining if a being is undergoing negatively valenced psychological states. This is far from easy. While we can easily recognise states like pain or fear in animals like dogs and cats, matters become far harder in other species. Note the difficulty involved, for example, in determining whether bees feel pain, or whether a trout kept alone in a tank feels distress at its isolation. Likewise, determining whether an animal's aversion to a stimulus is driven by, for example, instinctual or conditioned responses as opposed to negative psychological feelings or emotions is highly challenging. This is not to deny that a properly scientific account of valenced states may be possible, but no such framework currently exists, and there is still widespread disagreement even on many basic questions such as whether fish can feel pain (Key, 2015).

Establishing the occurrence of negatively valenced psychological states will be yet harder in the case of artificial beings, not least because artificial systems are unlikely to possess straightforward analogues of canonically unpleasant states like hunger, fear, or pain undergone by humans and animals. Some machines may – and indeed, already do – engage in forms of self-regulation with at least superficial similarities to these states; a robot might monitor its own power levels, for example, and spontaneously 'choose' to return to a charging station when its battery is low (Castro-González, Malfaz, & Salichs, 2013). But given the major differences in cognitive architecture between biological and artificial systems, we should be cautious of assimilating such states to the kinds of valenced psychological states undergone by humans and animals. Note also that many forms of self-regulation in humans and other animals – of the immune system, for example, or the production of stomach acids – operate at a subpersonal level below the level of awareness. Simply building robots capable of self-regulating various needs, then, will not demonstrate that they instantiate states of a kind that could give rise to experiences of suffering or felt unpleasantness.

The second challenge, which is perhaps more fundamental, concerns how we might attribute consciousness to a non-human entity. The very notion of suffering arguably presupposes such a capacity for conscious experience, at least in an everyday sense of the term (we do not normally think that individuals can suffer after they are dead, or when they are in comas, for example), yet there is no consensus on how we can reliably attribute consciousness to animals, let alone to artificial systems. Again, this is an area of intense interest within cognitive science, but expert opinions vary wildly; some philosophers and scientist extend consciousness to all vertebrates and some invertebrates (see, e.g., Barron & Klein, 2016), while others have suggested the whole question of non-human consciousness is

ill-posed (Carruthers, 2018).

The very possibility of conscious machines is a hugely controversial and complex issue, and has generated a copious literature in its own right, thus lies beyond the scope of the present paper. What I would stress in the present context, however, is that however great the challenges are for assigning consciousness to animals, the corresponding challenges for assigning it to machines will be far greater. In the case of animals, we can at least appeal to broad similarities in cognitive architecture and material composition, as well as a shared evolutionary history. By contrast, most artificial systems possess – and will continue to possess – wildly divergent forms of cognitive architecture and processing substrate from those found in human beings.

A last redoubt for those who wish to ground machine moral patiency in a capacity for suffering might be to appeal to verbal report. If a person sincerely reports that they are in pain, for example, we typically take regard the truth of the report as given. If a machine tells us it is suffering, then, should we not simply believe it? Needless to say, machines that can communicate with us fluently using natural language are, for now, a distant dream. However, even if such machines existed, it is not clear that their reports could be taken at face value. One famous argument to this effect is given by John Searle in his 'Chinese Room' example (Searle, 1980). In short, Searle asks us to imagine a scenario in which he – a non-Chinese speaker – is ensconced within a room, his only form of communication with the outside world being via a slot through which messages in Chinese can be passed. Within the room are a set of instructions in English for responding to messages posted through the slot. These instructions say nothing about the content of the messages, but merely specify via a vast database the appropriate characters to write in response to a given message, thus allowing Searle to carry out a semblance of a conversation with an external observer. Yet in such a situation, Searle suggests, neither he nor the broader system of which he is a part could be said to *understand* the messages in question.

The broader lesson Searle derives from this case is that there is more to understanding – and by extension, consciousness and mentality – than simply executing a program. For while Searle's example involves a human being, we can equally imagine a scenario in which the role played by Searle in the thought experiment is filled by a simple computer program that dutifully follows instructions for the production of symbols in response to inputs, yet likewise intuitively fails to understand Chinese.

It is easy to imagine how a similar thought experiment might constructed so as to create a system that ostensibly expresses and reports pain. Thus imagine that, in addition to the slot, the Chinese Room is equipped with a series of lights coupled to external sensors for detecting damage to the room. The occupant of the room might be instructed that, for example, when a specific red light goes on, they are to press a button and post a given character through the door, where this button activates a pained cry and the characters said 'ow, my foot!' or some other expression of pain. The occupant might similarly be tasked, when given a specific series of Chinese characters that (unbeknown to them) asked if they were in pain, to respond with a set of characters which (again, without their knowledge) stated 'yes, terrible pain'. Yet it seems clear that no suffering would be occurring in such a case.

The Chinese Room argument has been subject to many criticisms and attempts at rebuttal (e.g., Dennett, 2013). Even without endorsing Searle's broader argument, however, I

would suggest cases like the Chinese Room illustrate the fact that we can conceive of scenarios in which an artificial system produces at least many of the outward indicators of pain or suffering yet fails to undergo the relevant conscious states. It thus provides a further demonstration of the daunting challenge for the actual employment of notions like suffering as the grounds for artificial moral patiency.

3. Pathways to moral patiency (2): preferences

I would suggest, then, that attempts to ground artificial moral patiency in a capacity for suffering face severe obstacles in their practical employment. In light of these challenges, we might look for an account of moral patiency that did not explicitly rely on still poorly-understood notions like suffering or consciousness. One important alternative in this regard is the idea that a being might possess some limited form of moral patiency just in virtue of having *preferences*. This is a suggestion cogently developed by Marian Dawkins in regard to animal welfare. Dawkins begins with the observation that consciousness is still too poorly understood to ground a scientific approach to the welfare of non-humans. As she puts it, "we know so little about human consciousness… that we do not know what publicly observable events to look for in ourselves, let alone other species, to ascertain whether they are subjectively experiencing anything like our suffering" (Dawkins, 2008)

Instead, Dawkins suggests that we focus on two objectively assessable notions, namely animals' health and their preferences. These, she suggests, provide unambiguous indications of whether things are going well or poorly for an animal. If an animal is suffering from a lameness or a wasting disease, its welfare is being compromised. Similarly, if an animal has very strong preferences that are not being met – for food, mating, or a certain kind of environment, say – then their individual interests are not being satisfied. As she puts it, "'having what they want' is a shorthand way of covering a wide variety of way in which animals can, in a publicly observable way, show us whether the environment they are in is positive (something they like and want to continue with) or negative (something they dislike and want to escape from or avoid)" (Dawkins, 2017). Crucially for Dawkins' purposes, she believes these criteria provide a grounding for welfare and moral patiency even in the absence of evidence that the animal in question is conscious.[4]

Dawkins is primarily concerned with the practical question of measuring and improving animal welfare rather than the issue of what makes a being a moral patient to begin with. Nonetheless, her theory perhaps suggests a more practical way of grounding artificial moral patiency than the suffering. Rather than asking whether a system undergoes conscious suffering, we might instead ask whether it has robust preferences. Again, however, this faces a number of important objections, both as a general theory of non-human well-being and in its application to artificial beings in particular.

As an initial point, some may doubt whether any theory of moral patiency that is neutral on issues of consciousness and sentience is properly defensible. Intuitively, part of the reason we care about an animal is that its ill-health and the non-fulfilment of its strong preferences cause it to suffer.

---

[4] Dawkins' approach has much in common with desire theories of well-being, which take an individual's welfare to a function of whether their desires are satisfied (Griffin, 1986). There are many important historical objections to desire satisfaction views and variations of the basic position designed to answer them. See Heathwood (2006) for a review.

Even if we grant Dawkins' claim that moral patiency does not require consciousness, there are still difficulties with applying the theory to non-humans. For one, it is not clear how to make sense of the idea of health in relation to artificial systems. In the case of animal health, we can appeal to notions such biological function, evolutionary history and species norms.[5] Somewhat loosely, we might say that a dog with a lame leg is unhealthy because, first, the function of legs is to aid locomotion, and the dog's locomotion is impaired; second, because dogs evolved to have four functioning legs; and third, because, *ceteris paribus*, dogs do in fact have four functioning legs. By contrast, in the case of artificial beings, no such norms are available: robots or virtual beings are not members of well-defined species, nor can we ask if they are functioning in ways specifically selected for by past adaptive pressures. At best, we might appeal to the intentions of a designer: if a robot was designed to have two functioning legs, and only one is functioning, then we might say its interests are thereby harmed. Absent broader criteria for determining *which* machines to consider in this way, this proposal has a whiff of absurdity to it, however: we do not consider a robotic vacuum cleaner's welfare to be diminished when it has a broken wheel.

Dawkins' second basis for non-human moral patiency, namely preference satisfaction, might seem more readily applicable to artificial systems. We certainly attribute goal-directed agency to artificial beings, as when we say that, for example, that a robotic vacuum cleaner is attempting to skirt around an obstacle, or a non-player character in a videogame is trying to attack us. However, as suggested by these examples, the notion of agency in question is a very thin one indeed; few would suggest that vacuum cleaners or computer-controlled characters in their current state have any claims to moral patiency, even if we grant they have goals in some limited sense.

This prompts the question of how to determine whether a being has *morally relevant* preferences, and it is not one that is easily answered. Dawkins herself does not address the issue, and does not say, for example, whether we should lend any moral weight to the aversive responses of nematode worms or the appetitive responses of plants. Some philosophers in the preference utilitarian tradition (including Singer, 1979, and Varner, 2012) limit moral patiency to conscious or sentient beings with preferences, but this would of course be at odds with Dawkins' goal of excluding problematic notions like consciousness. Still, there are various other ways we might attempt to pin down a morally important notion of preference. One would be to identify the possession of true preferences with a capacity for negative and positive forms of reinforcement learning. On this view, we might say, for example, that a being has a preference for some state S if S can function as a reward, where this is understood as a state that reinforces behaviours that led to the obtaining of S. Again, however, this runs the risk of casting our net too broadly; transsected rat spinal cords that are completely cut off from the rat's brain can learn via negative reinforcement to raise the rat's legs (Liu et al., 2005). Moreover, reinforcement learning in various forms is already ubiquitous in artificial intelligence research, being present in systems such as AlphaGo, yet such programs are surely not plausible candidates for moral patiency.

It is of course possible that ongoing work in cognitive science may identify a principled criterion for distinguishing morally salient 'thick' preferences from simpler forms

---

[5] There is a rich debate in the philosophy of medicine concerning how best to analyse notions like health and disease. See, e.g., Kitcher (1996) for a more detailed discussion.

of appetitive and aversive behaviour and learning. However, absent such a criterion, a preference-based approach to artificial moral patiency will face challenges of application scarcely less daunting than those faced by a suffering-based approach.

4. Pathways to moral patiency (3): Personhood

A third strategy for grounding artificial moral patiency that is worth discussing is via the claim that some future artificial beings may constitute *moral persons*.[6] Intuitively, we associate personhood with the possession of a certain kind of moral status, and there are a number of distinct proposals for what marks out a person in the relevant sense (see, e.g., Frankfurt, 1971, Regan, 2004, and Taylor, 1985). For reasons of brevity, however, I will focus on the concept of personhood given by Kant (1798/2002), who claims that "The fact that the human being can have the representation "I" raises him infinitely above all the other beings on earth. By this he is a person….that is, a being altogether different in rank and dignity from things, such as irrational animals".

Specifically, Kant links personhood to the possession of rationality, that is, a capacity to reflect on and be moved by reasons for action, above all (for Kant) *moral* reasons. Thus contemplating a wallet left on the ground, I can recognise my moral duty as giving me a reason to turn it in to the police rather than pocket it. It is widely assumed that most non-human agents lack such reflective capacities, their actions instead being motivated by non-reflective states such as appetites and urges. As Korsgaard (1996) puts it, "A lower animal's attention is fixed on the world. Its perceptions are its beliefs and its desires are its will. It is engaged in conscious activities, but it is not conscious of them."

Let us now consider whether – and under what circumstances – an artificial being might satisfy this criterion for personhood.[7] This can be somewhat crudely rephrased as the question of whether an artificial being could act on the basis of reasons, and in particular, moral reasons. The idea of robots that act in accordance with moral rules may seem an odd one, but is a common theme in science fiction, particularly in the work of Isaac Asimov. Asimov proposes three (later four) "Laws of Robotics" (2004) that constrain robot behaviour. These require, first, that robots may not injure a human or allow a human to be injured; second, that robots must obey orders from humans, except where this conflicts with the first law; and third, that robots must seek to preserve their own existence except where this conflicts with the first and second laws. Asimov's large corpus of stories demonstrate robots acting in accordance with these laws, accidentally violating them, or resolving conflicts between them, and make vivid the idea of robots that engage in moral deliberation.

While Asimov's work is of course fiction, it seems entirely possible that we might equip artificial beings with rules for the determination of behaviour from which they were unable to deviate. Alternatively, we might attempt to make it *overwhelming rewarding* for a robot or artificial system to act in accordance with our moral rules. Discussion of such

---

[6] We should distinguish the notion of moral personhood from legal personhood. Note, for example, that legal personhood is granted to corporations and trade unions, as well as natural objects such as the Whanganui and Ganges rivers and Mount Taranaki. While such entities may have moral relevance in the sense discussed in Section 1, few would consider them to have intrinsic interests of the kind that ground moral patiency.

[7] Though I here present Kant's notion of personhood as separable from questions of consciousness, note that it is likely given Kant's broader philosophy of mind that he would consider the prerequisites for reflective thought to include self-awareness and consciousness.

constraints is a lively topic in debates about the future of superintelligence (e.g., Bostrom, 2017, Ch.13). However, from a Kantian perspective, we should be careful to distinguish between a system that superficially acts in accordance with a moral rule and one that acts rationally on *the basis* of moral reasons. A capacity for acting in accordance with moral principles for reasons of compulsion or pleasure would not suffice for personhood in Kant's terms (Kant, 1785/2002). What would instead be required for Kantian personhood would be a capacity to autonomously choose to act on moral reasons.

Needless to say, an artificial being capable of such autonomous deliberation lies far beyond the scope of present artificial intelligence, though perhaps not beyond the realms of possibility (see Bostrom, 2017, Ch.7). Note, however, that even if a superintelligence seemed to have developed ethical principles autonomously, a Kantian personhood framework to moral patiency would face a further challenge of interpretation, namely the question as whether the artificial system arrived at the principles via reason alone or through the complex responses to its reward function; the latter would surely not satisfy Kant's demanding conditions for autonomy. Moreover, to *rely* on Kantian personhood as a standard for moral patiency seems to be setting the bar for artificial moral patiency very high. Kant's own view, for example, has been frequently criticised for failing to extend moral patiency to animals (Korsgaard, 2009), and, as noted in Section 2 above, it plausible that we will build artificial systems that possess a strong intuitive case for moral patiency on grounds of cognitive complexity long before we achieve human-level AI. For all its intellectual interest, then, I would suggest we have good grounds for looking for alternatives to the Kantian personhood approach sketched here.

5. A Biological Analogy framework for artificial moral patiency

I have suggested that all three routes to artificial moral personhood face serious obstacles and limitations. A suffering-based approach faces daunting challenges in its application, most notably with regards to determining whether an artificial system is conscious. A narrowly preference-based approach faces the difficulty of pinning down a notion of preference that does not lead us to spuriously include nematode worms and robotic vacuum cleaners as moral agents. Finally, a personhood-based approach faces the worry that it is overly restrictive, excluding beings with genuine interests simply on the grounds that they lack the capacity for reflective moral attitudes.

In this final section of the paper, I wish to consider a fourth account that I term the *Biological Analogy* approach. In short, this claims that we should (tentatively) ascribe moral patiency to an artificial system once we judge it to possess cognitive and behavioural capacities broadly equivalent to those of a non-human animal to which we are already reflectively inclined to ascribe moral patiency. To put it more simply, if we judge that dogs are moral patients, and we build an artificial system whose cognitive and behavioural capacities are broadly equivalent to those of dogs, then we should similarly ascribe moral patiency to that system.

One immediate advantage of this approach is that it allows us to avoid certain risks of 'biocentrism' in our ascription of moral patiency, that is, the irrational prejudicing of biological over artificial entities. We naturally – though questionably – attribute states like pain and distress more readily to animals such as dogs and cats whose characteristic behavioural responses to stimuli superficially mirror our own than to more ecologically alien

creatures such as octopuses or fish. Likewise, there is a danger that we will be irrationally more inclined to attribute moral patiency to biological than to artificial entities in virtue of superficial similarities of appearance or behaviour. Via coupling our protections for artificial entities to our judgments about animal moral patiency across carefully constructed cognitive and behavioural parameters, we are less vulnerable to the risk of discounting compelling evidence of moral patiency on the basis of superficial differences between biological and artificial systems.[8]

The Biological Analogy framework as sketched above needs to be fleshed out to answer some obvious worries. The first is the issue of whose opinion matters in determining whether a given analogical organism is a moral patient to begin with. As noted above, there is still considerable disagreement within cognitive science about which animals are moral patients, as well as major discrepancies across different legal systems concerning which animals have even minimal protections. Nonetheless, as long as we keep in mind that our ascriptions of moral patiency to machines made under the Biological Analogy framework will be tentative and fallible, I would suggest the difficulties here are not insuperable. Policy-makers already face the challenge of determining which animals qualify as moral patients, and draw upon a wide range of resources to do so, taking into account both popular opinion and that of philosophers and scientists. As a preliminary step, then, different jurisdictions might use existing frameworks for animal rights within their specific legal system as the basis for fixing the behavioural and cognitive capacities required for an artificial system to qualify for moral patiency.

A second concern for the Biological Analogy account is *which* behavioural and cognitive capacities to include as relevant. While we might have little difficulty in dismissing the relevance of defecation or perspiration for moral patiency, other capacities are harder to gauge. Should things such as mating drives, hunger, or social behaviour qualify as relevant? Again, while this may seem a daunting question, I would suggest that in practice such difficulties can be overcome via careful experimentation and reflection. Imagine, for example, that we believe chickens to be moral patients. We can then ask which of their capacities, if absent in a specific chicken, would lead us to reasonably judge them not to be a moral patient. If a brain-damaged chicken were not to have mating instincts, for example, or lacked normal drives for hunger and thirst but was otherwise behaviourally identical to its conspecifics, we would surely not consider its moral status to be significantly different from that of other chickens. If, by contrast, it did not engage in any kind of social behaviour, did not display purposive agency, or had a dramatically diminished capacity for learning, we might reconsider the matter.

We could also use the study of the range of behaviours available to decapitated or decerebrated animals (who are unlikely to be conscious or otherwise qualify as moral patients) to constrain the relevant cognitive and behavioural parameters of comparison. For example, it was demonstrated by Ferrier (1886) that a decapitated frog will still rub its thigh

---

[8] There is, of course, a risk that existing biases towards or against animal species will be carried over to their artificial equivalents; a robot whose cognitive and behavioural capacities closely resembled those of a ‘charismatic’ animal such as a dog, for example, might be unfairly judged to be a better candidate for moral patiency than one whose capacities resembled those of a reptile or fish. The best answer to this worry, I would suggest, is to use reflective scientific methods to first address biases in the legal protections afforded to different animal species, and carry over such corrections to our treatment of artificial moral patients.

in response to the application of a drop of acetic acid, while Gentle (1997) showed that decerebrate chickens would still modulate pecking behaviour in response to an injured beak if the injury preceded the removal of their forebrain, suggesting such behaviours are accomplished by relatively low-level mechanisms. Via a combination of empirical and reflective methods such as these, we might reasonably attempt to establish a set of cognitive dimensions relevant for moral patiency.

A third reasonable worry might be that some (perhaps highly intelligent) artificial systems would 'fall between the cracks' of the Biological Analogy framework simply insofar as they lacked appropriate biological analogues. A virtual agent with considerable cognitive powers yet lacking any sort of body (such as the famous HAL 9000 of *2001: A Space Odyssey*) would simply be too different from any real world organisms for reasonable comparison. In response to this, I would note the Biological Analogy framework should be understood as (at least in some respects) a deliberately conservative proposal, to be buttressed by other approaches; the fact that a given artificial system fell outside its boundaries would not automatically guarantee that the system lacked all moral status.

6. Why machines aren't moral patients (yet)

The Biological Analogy framework seems to me to constitute a practical and commonsense method for establishing artificial moral patiency. Nonetheless, I recognise that some will balk at the idea that any near-term artificial system might have the same claim to ethical consideration as a cat or dog. As mentioned in Section 1, moral patiency need not entail moral obligation. However, I am nonetheless in broad agreement that artificial moral patients remain a distant prospect, as I will now briefly discuss. In short, this is because even a cursory application of the Biological Analogy framework to current artificial systems illustrates their dramatic cognitive shortcomings in comparison to animals, even including simpler organisms such as insects and crustaceans whose status as moral patients is far from clear.

Above all, I would note that there is a glaring difference in cognitive capabilities between animals and machines in respect of *General Intelligence*; that is, the capacity of an entity to reliably and flexibly manage and accomplish its goals. By this metric, many animals perform extremely impressively, as demonstrated by the fact that most can accomplish feeding, mating, navigation, and other biological functions in a wide range of environments and conditions. The Western honey bee, for example, inhabits every continent except Antarctica, and has adapted its behaviour to reflect different habits. Some of the adaptations may involve region-specific genetic selection, but others – such as synchronizing of colony cycles to coincide with local flora, or the adoption of specialised foraging techniques in desert environments – are likely to be learned responses to varying habitats (Whitfield et al., 2006). The migratory patterns of birds likewise offer a striking demonstration of the capacity of animals to navigate across and thrive in different environments. The Great snipe, for example, spends its summers in the cool conditions of Russia and north-western Europe, but migrates some 5500km in winter to the much warmer environments of sub-Saharan Africa, making stop-offs in southern Europe and north Africa along the way.

By contrast, the performance of current artificial systems is generally extremely brittle. To give a simple example, it is rare to see animals lose control of basic motor functions, falling over or getting stuck, yet as any robotics researcher (or owner of a robot

vacuum cleaner) can attest, this is a perennial problem for artificial systems. Outside of tightly-regulated training environments, artificial systems are vulnerable to a snowball effect of small errors, and can often be easily tricked into making spectacular errors, as demonstrated by the case of 'adversarial examples' in which machine vision systems make bizarre errors when presented with subtly perturbed images (see Fig.1 below).

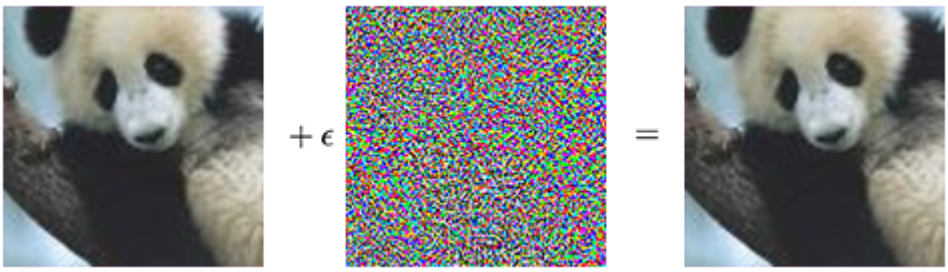

Fig. 1. The GoogLeNet image classifier correctly categorises the first image as a panda, but after the application of an adversarial filter, identifies it as a gibbon. (Goodfellow, Shlens, & Szegedy, 2014)

In respect of flexibility and learning, biological systems display a striking ability to acquire new skills and apply existing skillsets to new tasks. Bumblebees, for example, are capable of rapidly learning to perform novel tasks such as rolling a ball into a hole via watching conspecifics perform the same task (Loukola, Perry, Coscos, & Chittka, 2017). Among birds, there are numerous demonstrations of impressive flexible causal reasoning with novel tools (Jelbert, Taylor, Cheke, Clayton, & Gray, 2014), and a rich body of data shows a capacity to flexibly adapt caching behaviour to accommodate different foodstuffs and retrieval opportunities (Clayton, Bussey, & Dickinson, 2003).

This is again in marked contrast to artificial systems. The challenge of building machine intelligences capable of transferring knowledge across different tasks is one of the great unsolved problems of current AI research, and while some progress has been made, even cutting edge artificial systems display only minor performance improvements when transferring knowledge between closely related tasks (Lake, Ullman, Tenenbaum, & Gershman, 2016). A related long-standing difficulty encountered by machine learning researchers is the tendency of artificial systems trained to perform one task to suffer 'catastrophic forgetting' when assigned to another one. While some progress has been made in this domain, the prospect of building systems that can engage in the same kinds of rapid and effortless task-switching readily accomplished by even simple animals remains a distant one. Finally, note that most current machine learning systems require comparatively vast sets of training data or extensive practice time to achieve high levels of performance, again a dramatic contrast with biological intelligence (consider that a newborn fawn learns to stand up after 10 minutes and walk smoothly in just 7 hours).

While general intelligence of the kind discussed here reflects just one suite of cognitive abilities, it is arguably serves as a key component in a huge range of high-level capacities, including many that we might be tempted to link to moral patiency, such as awareness, autonomy, robust goal directedness, and understanding. Within the Biological Analogy framework, then, the kinds of abilities encompassed by the broad notion of general intelligence are likely to among those which we prioritise for the purposes of establishing

relevant cognitive similarities between biological moral patients and candidate artificial moral patients. If this is correct, then artificial moral patients with the same claim to moral patiency as dogs, birds, or even honeybees are likely to be a long way off.

3. 7. Conclusion: why artificial moral patiency still matters

My goal in this paper has been to explore the possibility of artificial moral patiency, and to suggest how we might come to reasonably identify artificial systems as moral patients. I began by explicating the notion of moral patiency, before considering three well-established bases for moral patiency in humans and animals, namely a capacity for suffering, the possession of preferences, and personhood. Identifying limitations with all pathways, I suggested an alternative pragmatic approach I termed the Biological Analogy framework, while noting that no current artificial system come close to qualify as persons by its lights.

Despite the fact that artificial moral patients are still a distant target, there is nonetheless great value in reflecting now about their possibility. Artificial moral patients could proliferate rapidly, especially if realised in virtual environments; by virtue of their relative cognitive inscrutability compared even to animals such as fish and octopuses, they might be hard to identify when they do arise; and it might be comparatively easy to avoid causing harm to them via small alterations of design, as compared to the large changes in lifestyle and farming necessary to eliminate or minimise harm to non-human animals. Most of all, however, I would suggest that close attention to the conditions and possibility of artificial moral patients may serve as a further spur to the development of broader tools and frameworks for assessing welfare and moral status in the broader domain of all non-human forms of life.

REFERENCES

Asimov, I. (2004). *I, robot* (Vol. 1). Spectra.

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences Proc Natl Acad Sci USA*, *113*, 4900–4908.

Bostrom, N. (2017). *Superintelligence*. Dunod.

Carruthers, P. (2018). Comparative psychology without consciousness. *Consciousness and Cognition*, *63*, 47–60.

Castro-González, Á., Malfaz, M., & Salichs, M. A. (2013). An autonomous social robot in fear. *IEEE Transactions on Autonomous Mental Development*, *5*(2), 135–151.

Clayton, N. S., Bussey, T. J., & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews. Neuroscience*, *4*(8), 685–691. https://doi.org/10.1038/nrn1180

Dawkins, M. S. (2017). Animal welfare with and without consciousness. *Journal of Zoology*, *301*(1), 1–10. https://doi.org/10.1111/jzo.12434

Dawkins, Marian Stamp. (2008). The Science of Animal Suffering. *Ethology*, *114*(10), 937–945. https://doi.org/10.1111/j.1439-0310.2008.01557.x

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. WW Norton & Company.

Ferrier, D. (1886). *The functions of the brain*. Smith, Elder.

Frankfurt, H. G. (1971). Freedom of the Will and the Concept of a Person. *Journal of Philosophy*, *68*(1), 5–20. https://doi.org/10.2307/2024717

Gentle, M. J. (1997). Pain-related behaviour following sodium urate arthritis is expressed in decerebrate chickens. *Physiology & Behavior*, *62*(3), 581–584.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *ArXiv:1412.6572 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1412.6572

Griffin, J. (1986). *Well-Being: Its Meaning, Measurement and Moral Importance*. Clarendon Press.

Heathwood, C. (2006). Desire satisfactionism and hedonism. *Philosophical Studies*, *128*(3), 539–563.

Jaworska, A., & Tannenbaum, J. (2018). The Grounds of Moral Status. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2018). Retrieved from https://plato.stanford.edu/archives/spr2018/entries/grounds-moral-status/

Jelbert, S. A., Taylor, A. H., Cheke, L. G., Clayton, N. S., & Gray, R. D. (2014). Using the Aesop's Fable Paradigm to Investigate Causal Understanding of Water Displacement by New Caledonian Crows. *PLOS ONE*, *9*(3), e92895. https://doi.org/10.1371/journal.pone.0092895

Kamm, F. M. (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford, New York: Oxford University Press.

Kant, I. (2002). *Groundwork for the Metaphysics of Morals*. Oxford University Press.

Kant, I. (2007). Anthropology From a Pragmatic Point of View (1798). In I. Kant (Ed.), *Problemos* (pp. 177–198). Cambridge University Press.

Key, B. (2015). Fish do not feel pain and its implications for understanding phenomenal consciousness. *Biology & Philosophy*, *30*(2), 149–165. https://doi.org/10.1007/s10539-014-9469-4

Korsgaard, C. (2009). Facing the Animal You See in the Mirror. *The Harvard Review of Philosophy*, *16*(1), 4–9. https://doi.org/10.5840/harvardreview20091611

Korsgaard, C. M. (1983). Two Distinctions in Goodness. *The Philosophical Review*, *92*(2), 169–195. https://doi.org/10.2307/2184924

Korsgaard, C. M. (1996). *The Sources of Normativity* (Vol. 110). Cambridge University Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *ArXiv:1604.00289 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1604.00289

Liu, G. T., Ferguson, A. R., Crown, E. D., Bopp, A. C., Miranda, R. C., & Grau, J. W. (2005). Instrumental Learning Within the Rat Spinal Cord: Localization of the Essential Neural Circuit. *Behavioral Neuroscience*, *119*(2), 538–547.

Loukola, O. J., Perry, C. J., Coscos, L., & Chittka, L. (2017). Bumblebees show cognitive flexibility by improving on an observed complex behavior. *Science (New York, N.Y.)*, *355*(6327), 833–836. https://doi.org/10.1126/science.aag2360

McMahan, J., & McMahan, A. P. of P. and R. S. J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford University Press.

Philip, K. (1996). *The lives to come: The genetic revolution and human possibilities*. London: Penguin.

Rachels, J. (1991). *Created from Animals: The Moral Implications of Darwinism*. Oxford, New York: Oxford University Press.

Regan, T. (2004). *The Case for Animal Rights*. University of California Press.

Scanlon, T., & Scanlon, A. P. of N. R. M. P. and C. P. T. M. (1998). *What We Owe to Each Other*. Harvard University Press.

Searle, J. R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*(3), 417–57.

Singer, P. (1979). *Practical Ethics*. Cambridge University Press.

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014). Defining and assessing animal pain. *Animal Behaviour*, *97*, 201–212. https://doi.org/10.1016/j.anbehav.2014.09.007

Steinbock, B. (2011). *Life Before Birth: The Moral and Legal Status of Embryos and Fetuses, Second Edition*. Oxford University Press.

Taylor, C. (1985, March). The concept of a person. https://doi.org/10.1017/CBO9781139173483.005

Varner, G. E. (2012). *Personhood, Ethics, and Animal Cognition: Situating Animals in Hare's Two Level Utilitarianism*. Oup Usa.

Whitfield, C. W., Behura, S. K., Berlocher, S. H., Clark, A. G., Johnston, J. S., Sheppard, W. S., … Tsutsui, N. D. (2006). Thrice out of Africa: ancient and recent expansions of the honey bee, Apis mellifera. *Science (New York, N.Y.)*, *314*(5799), 642–645. https://doi.org/10.1126/science.1132772