Suffering Subroutines: On the Humanity of Making a Computer that Feels Pain

Meghan Winsby
Western University, London ON
mwinsby@uwo.ca

Abstract

In this paper I question the moral permissibility of developing sentient machines as part of an artificial intelligence (AI) research program. Initially, and from the basic assumptions that this is possible, that pain has a certain (unpleasant) character, and that beings that can feel pain are owed some level of moral consideration, I argue that pain engineering in AI is prima facie morally wrong. I then consider some ways in which proponents may object to—or at least dampen—this initial position.

1 Introduction

Traditionally, one of the more "peripheral"[1] tasks assigned to moral philosophers has been that of determining the scope of moral concern; that is, determining the sorts of things that have moral standing. In particular, the question of which non-human entities are entitled to moral consideration, if not full respect, has only recently come into its own as an independent enterprise. The moral standing of animals, discrete parts of the natural environment, ecosystems, or nature as a whole has been of primary concern to animal-interest advocates and environmental philosophers. In these and other areas of applied ethics, strict anthropocentric views are enjoying ever more limited support. Non-human animals, in particular, have been widely recognized as possessing varying levels of moral *patiency*.[2] The standing afforded to a being that lacks moral agency, but nonetheless counts as moral patient, allows for entitlement to at least a minimal degree of moral consideration.

Discussions surrounding the moral status of machines have an even shorter history, as does machine ethics more broadly. The vision of a world in which human artifacts are possessed of thoughts, feelings, rights and responsibilities may seem the stuff of science fiction fantasy. However we live in an age of automated weapons systems, military automaton research, and empathy-simulating robots. Questions about displaced responsibility and blame, [3] and human emotional responses and attachments to machines are already driving serious conversations. Though we have not yet made the technological leap from simulation to artificial intelligence (AI) and/or artificial consciousness, anticipation of a morally more complex set of relations between humans and machines appears warranted.

[1] Martin Schönfeld (1992) 353.

[2] Broadly, moral patiency recognizes beings capable of being benefited or harmed, and this fact is sufficient, on some accounts, for moral standing for those beings.

[3] Robert Sparrow, for example, argues that the advent of military robots capable of acting autonomously could spell the fragmenting of the chain of responsibility for operations that may be classified as war crimes.

To this end, investigations into the moral status of machines—attempted answers to the ―"machine question"[4]—have focused on machines as loci of blame and responsibility, bearers of rights and entitlements, and possessors of moral agency and/or patiency. In what follows I want to pursue a related, but somewhat different question. In this essay, I will ask whether the *project* of developing sentient machines is *itself* a humane one. In particular, I want to draw attention to the moral questionability of a research program concerned with generating pain experience. My aim here is not to come down cleanly on any side of the problem, but rather to set it up starkly and suggest some possible avenues for further discussion.

2 Three Assumptions

First, I will assume that the apparent challenges associated with developing AI capable of sentience (the ability to experience pain and pleasure) and self-awareness can be, in principle, overcome. This assumption will be unpopular with those who believe that pain and other mental states are singularly realizable,[5] and even on a functionalist framework, the prospect of machines that feel pain is not without its critics. For those who think a pain-feeling entity of human-made material conceptually untenable, this essay is not for you. For those of us still on board (which is likely most of us),[6] this will also involve setting aside epistemological questions like *even if we* could *develop a machine that experiences pain, how could we* know *we have achieved our objective?* Though these worries are not uninteresting, the problem of the knowability of pain experience in other minds is one that extends beyond the scope of my argument. I think it safe to assume that an autonomously functioning machine that is capable of experiencing pleasures and pains, and is able to recognize itself as subject, represents a major benchmark—maybe even the endpoint—of perhaps the more idealistic AI research projects. Where AI here is understood as a cognitive scientific undertaking[7] aimed at advancing our understanding of the operations of the human mind, this seems a worthy goal, however fantastic. So let's assume that it is *possible* to realize pain experience in a machine or some other artifact.

Second, I am going to assume a view of pain according to which a pain experience has two components. One is a descriptive element that tells the subject where, when and to what magnitude the pain occurs: it locates the pain. The second is an affective component. Pains are felt by the subject to be bad, or unpleasant. These two components are individually necessary and jointly sufficient for a complete pain experience. Important to note is that this view of pain is functionalist, so that pains are system states with a certain functional role. The functional role of a state involves its typical causal connections to sensory inputs, behavioural outputs, and other states.[8] The 'pain role' includes a disposition to be caused by bodily damage, to cause avoidance behaviour (among others), and to elicit judgments that something has gone wrong with the subject's body, for example. It may be that pain plays a crucial, if not *ineliminable* functional role in an entity's cognitive system, despite its necessarily carrying a negative affective component for the subject.

In light of these first two assumptions, the determined AI researcher might be encouraged by the possibility of creating machines capable of instantiating more than mere simulations.[9] She may be

[4] See especially David J. Gunkel (2012).

[5] I have in mind John Searle, for one, whose biological essentialism requires human wetware in order to realize *understanding* (his term).

[6] The idea that consciousness is multiply realizable (so that it is in principle possible to instantiate cognitive systems in silicone, e.g.) is not overly controversial.

[7] As opposed to AI as *industry*—in the development of software, video games, search engines, manufacturing, etc.

[8] See Block, Ned (1980), esp. pg. 3.

[9] The sort of "simulating" I have in mind could be merely having the behavioural dispositions characteristic of pain, without instantiating a state that plays the rest of the pain-role.

understandably excited by the prospect of interacting with entities to which—or, with whom—we may relate in a more meaningful way, and the mutual understanding of pain experience can be counted as one (perhaps the most)[10] important stepping stone toward the realization of machine-human *empathy*. In humans (and many non-human animals), empathy is essential to successful social interaction. If human-machine social interaction is to advance, an important desideratum for AI research is surely the realization of empathetic machines.

These assumptions spell good news for human cognitive/neurological research as well. After all, a machine that can feel pain offers a new and richer avenue by which we can study pain—and the effective management thereof, for example—in *humans*. Treatments for hyperalgesia, complex regional pain syndrome, and other chronic pain could be carried out and tested using an artificial pain pathway generating genuine pain experience. This brings me to my third assumption.

Third, for the time being, I will assume a weak *sentientism* about membership in the moral community. Moral standing in this community entails that an entity‘s “continued existence and well-being or integrity are ethically desirable, and [he/she/its] interests in them carry positive moral weight.”[11] These interests impose corresponding duties on moral agents, and these duties are *direct*. On this view, sentience is sufficient for at least a minimal level of moral concern, such that we think it morally wrong to cause pain in beings that can suffer if we can help it, all things being equal. We have a direct duty not to engage in the inhumane treatment of entities that can be said to suffer pain. Sentience, then, is a fairly minimal feature that we use to distinguish non-moral entities from those to which we ascribe some moral patiency. I say “weak” sentientism because I want to emphasize that sentience need not be sufficient for *full* moral standing. That is, where the interests of a merely sentient being conflict with a sentient, self-aware and fully rational one, the interests of the latter will (or may, at least) outweigh those of the former. Moral standing, then, need not afford sentient beings consideration *equally* on this view. Each being may have moral standing, but its interests may not be given the same weight in moral deliberation as others' interests.

In sum, our assumptions at the outset are (1) that it is possible, and at some point down the line likely, that we can create a pain-experiencing machine; (2) that pain plays an important functional role in an entity‘s cognitive system, and that it necessarily involves unpleasantness for the subject, and (3) the ability to experience pain is typically grounds for moral consideration. In this regard—as I will elaborate below—it is normally our aim to *end it*, or to *avoid causing it*.

3 The Problem

Taken together, the above assumptions give us reason to pause. They ought to lead us to consider, for one thing, whether the creation of artificial moral patients is itself a morally legitimate pursuit. This is one question: is the creation of sentient beings good for *them*? Perhaps this pursuit is permissible, and it is morally unproblematic to multiply the pool of beings to whom we owe moral consideration (isn‘t this just what we do when we procreate?).[12] But what about the road we take to get there in the case of AI? This is a slightly different question. Given the realizability of pain experience in artificial substance, and that pain cannot be merely descriptive but also *unpleasant* for the subject, and that the ability to experience pain is grounds for humane treatment, the question becomes whether it is *humane* to design the software or engineer the hardware with which to implement a pain pathway. Is it permissible for an AI pain-engineer to bring pain into existence—in a sense, from nothing? What if this activity is undertaken in the service of some further end (either for them, or for human beings)?

[10] See Singer, Tania et al. (2004).

[11] Schönfeld (1992), 356.

[12] Here we may draw an analogy with anti-natalist arguments in the literature on procreative ethics.

Taking seriously the goal of achieving genuine empathy in machines, we have good reason to think that bringing about not only the capacity for pain, but its realization in artificial subjects is a necessary step in developing this disposition. In humans, the development of the perspective-taking stance involved in empathy appears to draw on cognitive resources obtained through past experiences of emotional and sensory feelings for the subject herself. Empathy refers to the ability to feel what others feel, and "accordingly, empathetic experience enables us to understand what it feels like when someone else experiences sadness or happiness, and also pain, touch, or tickling."[13] Additionally, there is evidence to suggest that the affective (unpleasant) component of pain is necessary, and in fact the only component actively engaged when empathizing with others' pain experiences. Data from one study suggest that we use "decoupled representations to understand the feelings of others, and that our ability to empathize has evolved from a system of representing our internal bodily states *and* subjective feeling states."[14]

If we have inklings that intentionally programming and implementing pain experience is morally shaky, then we ought to investigate this intuition.

4 The Principle of *Do No Harm*

On the face of it, there seems to be a straightforward answer to the question of whether intentionally implementing pain experiences is morally permissible. Even pre-theoretically, it is a widely held moral principle that we ought not to act in such a way as to cause harm. Moral rules in general, it can be argued, seek to minimize or avoid harm by limiting actions that cause or pose a great risk of causing harm. On a sentientist view of moral standing, those to whom we extend the courtesy of not inflicting harm are those able to experience pain and/or pleasure, as it is in this way that these beings stand to be harmed. Near the top of the list of paradigmatic harms is the infliction of pain, though it is also standard to characterize harm as the frustration of a patient‘s *interests*—specifically interests in her own welfare, among others. Even on this view, it is fair to say the experience of pain ranks highly among the specific welfare interests of moral patients. Sentient patients may possess a minimal set of interests (perhaps containing *only* an interest in not experiencing pain) dictating a duty for moral agents to avoid frustrating those interests,[15] all things being equal.

There are also the further questions about whether the AI pain-engineer causes harm to a sentient machine directly or indirectly, intentionally or unintentionally. If the mechanism via which an artificial being feels pain arises—unintentionally—out of the engineering of *other* cognitive processes, then we might say the action on the part of the AI pain-engineer is *not* one which causes harm directly. On a classical AI model, it is hard to see how the pain-engineer‘s actions could be construed as the indirect, and/or unintentional, cause of the pain. This is because on the classical model, each subsystem—including that which would implement the function of pain experience—would have to be individually programmed. There is little hope, then, of getting around pain experience that is directly and intentionally designed on the classical model, because developing such a system plausibly requires the testing of these systems. Thus, the classical model AI researcher is intentionally brining about pain experience.

Perhaps on a connectionist model—according to which consciousness emerges out of the complex interconnections between nodes within network—there is a way to deflect intention of this sort? Here it seems worse. At least on a classical model, we could imagine a near-perfect programmer, who need never test her algorithm. It is then in principle possible to engineer the capacity within a cognitive

[13] Singer et al. (2004) 1157.

[14] *Ibid*., 1161 (emphasis mine).

[15] Joel Feinberg's account of harm, for example, views them as setbacks to interests.

system, without *actually* causing a pain experience during its design. On a connectionist model, however, a network needs to be *trained*. Training a network to feel pain would involve giving the network a specific data set to deliver the intended behaviours and connections associated with pain experience. This would be repeated until the desired outputs (behaviours, connections to other states, etc.) are achieved. In essence, the pain experience—or at least its rudiments—would be brought about time and again until the machine is able to reliably exhibit the desired outputs (i.e. genuine pain experience). It is less plausible to imagine a programmer on this model who just happens to divine the correct configuration of nodes and the weightings of connections between them, so that the capacity for pain arises in the design, unrealized.

Whether direct or indirect, the programming of a pain subroutine or the training of a pain network seems to constitute a harm that is prima facie morally wrong, insofar as this activity intentionally brings into being not only the new potential or capacity for pain, but also actual instances of pain—both the subject *and* the experience. However, there are also various considerations that would count in favor of the permissibility of implementing pain experience as part of a comprehensive AI research program, *despite* this apparent harm. I will consider two below.

5 Objections

In light of the potential *good* to be gained from sentient AI (in fact, some may argue that sentience is itself *essential* to full-fledged AI), there are several ways to build a case for the permissibility of AI pain research.

Objection One: As we move forward, incorporating machines ever more seamlessly into day-to-day life, it will become vital that we have AI working alongside human beings that is equipped with *genuine* empathy, and not mere *simulations* of empathy. We want machines that do more than merely evince care and compassion through outward behaviour. We want to be able to *trust* that the machine itself has an investment in these relations with us. From what we know about empathy in humans, it is plausible to suppose that the only way to foster genuine empathy in an artificial network is by training that network to experience pain for itself. The importance of achieving empathy in AI cannot be over stated, the objection may go, as the AI *industry*—and the development of robotic caregivers[16] in particular—may very soon bring AI technology into our homes. Further, genuine empathy with human beings is arguably good for *them*, and this benefit to future conscious AI outweighs the harms attached to the pain suffered during these (early) stages of their development. A balancing of harms against potential benefits will always be a complicated endeavor, and the particulars of this evaluation are too numerous to list here.

Objection Two: In a different vein, there may be room for appeal to the doctrine of double effect in the case of AI pain research. It may be objected that the same worries about the permissibility of bringing pain-experiencers into existence can be run with respect to the morality of human procreation. Procreation, it is claimed, involves creating individual subjects of experience, capable of feeling pain and almost certainly *will* feel pain—even extreme pain—throughout the course of their lives. David Benatar, for example, argues that the badness of pain is enough to render existence itself a harm to sentient beings, thus making procreation morally wrong. This seems structurally analogous to the creation and development of sentient AI, and insofar as we are morally okay with human procreation, we ought to be okay with sentient AI creation. The reason, we might add, that we believe having children to be permissible—despite the fact that one‘s child or children will suffer—is that our action (procreation) is not undertaken with the intention to cause that suffering. We may view this suffering as a foreseen but unintended consequence of our action. Our intention is to bring into

[16] Not to mention companions of other sorts—pets, “sexbots,” etc.

existence the subject of a comprehensive set of experiences, of which pain is one, but the set is one we think worthwhile overall.

However there is perhaps a relevant *disanalogy* between human procreation and the creation of sentient AI. In light of what I have argued above, the development of sentient AI plausibly involves the *direct* and *intentional* implementation of *actual* pain experience. This is not so in the case of human procreation, however. Our intention here may reasonably be construed as one concerned with creating a life—complete with a wide range of experiences—and the subsequent bringing into existence of a pain-experiencer is merely a side-effect, although arguably a foreseen side-effect. Bringing a fully formed sentient being into the world involves intentionally generating the capacity for more pain experiences, whereas I have argued the development of sentient AI intentionally generates actual pain experiences.

There may be a way for the doctrine of double effect to get a foot in, however, as the AI pain-engineer may see herself as implementing pain experience—one of a number of cognitive functions—simply as a means to the more comprehensive end of functions necessary for achieving full-fledged sentient AI. It is also possible that the development of AI involves designing more general cognitive functions, out of which pain systems naturally "emerge," although at this point this approach is not well specified in AI research.

6 Conclusion: Propagating the Problem of Evil?

It would be unfortunate, if not tragic, if future AI came to justifiably resent their makers, and there is a sense in which the worries expressed in this essay echo a kind theodicy. How can it be compatible with our basic moral tenet of *do no harm* to intentionally bring about the experience of pain (the pain function in sentient AI)? The proponent of the moral permissibility of creating sentient AI might argue that pain experience is necessary to the development of the general functioning of a cognitive system. Arguments that try to justify this endeavor are reminiscent of those put forward in the attempt to reconcile the omniscience, omnipotence and omnibenevolence of God with the existence of evil in the world. One might argue that creating a cognitive system that can feel pain will have a better life overall—it is for the best that they feel pain. Similarly, the theologian might argue that the presence of evil is in fact to our overall betterment.[17]

However there are of course important differences, one of which is this: we are not—or do not purport to be—omniscient, omnipotent and omnibenevolent, and we recognize or ought to recognize our own epistemic, practical, and moral limitations. Whether this humility itself serves as a decisive reason to refrain from the development of sentient AI, or merely calls for extreme caution remains to be seen, and certainly more work ought to be done.

The exercise of restraint, with respect to AI pain research, may relevantly parallel an already widely accepted practice. Animal testing with *mild* animal suffering is a practice to which we may look for guidance, though this practice is of course not free from controversy itself. In the United Kingdom, for example, tests which involve the subjects suffering may be carried out—and the suffering classified as mild—so long as all measures are taken to minimize the level of pain during the course of these tests, and the proposed procedures to ensure this occurs satisfy a licensing board. It stands to reason that an AI pain engineer's chosen methodology could satisfy such a regulatory board, and the design and testing of artificial pain networks carried out with only *mild* suffering. In light of the knowledge and other benefits to be gained from the development of artificial pain networks, so long as the harm is minimized, the engineering of pain in an artificial being may be morally permissible. Of course, this will not satisfy a strict sentientist, who may still object that *any* amount of

[17] John Hick, for example argues that the existence of evil is necessary for the moral development of our souls, in preparation for the afterlife.

pain brought about when not absolutely necessary is too much pain, and unjustifiable in the name of AI research.

References

Block, Ned. 1980. "Functionalism." In Ned Block (ed.), *Readings in the Philosophy of Psychology*.

Dennett, Daniel. 1981. "Why You Can't Make a Computer that Feels Pain." In *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press. 190 229.

Gert, Bernard. 2004. *Common Morality: Deciding What to Do.* Oxford: Oxford University Press.

Gunkel, David J. 2012. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* Cambridge: MIT Press.

Hick, John. 1966. *Evil and the God of Love*. Macmillan.

Putnam, Hilary. 1975. "Brains and Behavior." In *Mind, Language and Reality: Philosophical Papers, Volume 2.* Cambridge, MA: Cambridge University Press. 3

Searle, John. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences*. 3: 417–57.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science,* 303(5661), 1157-1162.

Shönfeld, Martin. 1992. "Who or What has Moral Standing?" *American Philosophical Quarterly.* 29.4. 353-362.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy.* 24.1. 62-77.