AAAI: an Argument Against Artificial Intelligence

Sander Beckers

Utrecht University, Department of Philosophy and Religious Studies, Janskerkhof 13, 3512 BL Utrecht, the Netherlands,
srekcebrednas@gmail.com,
WWW home page: https://sanderbeckers.com

Abstract. The ethical concerns regarding the successful development of an Artificial Intelligence have received a lot of attention lately. The idea is that even if we have good reason to believe that it is very unlikely, the mere possibility of an AI causing extreme human suffering is important enough to warrant serious consideration. Others look at this problem from the opposite perspective, namely that of the AI itself. Here the idea is that even if we have good reason to believe that it is very unlikely, the mere possibility of humanity causing extreme suffering to an AI is important enough to warrant serious consideration. This paper starts from the observation that both concerns rely on problematic philosophical assumptions. Rather than tackling these assumptions directly, it proceeds to present an argument that if one takes these assumptions seriously, then one has a moral obligation to advocate for a ban on the development of a conscious AI.

1 Introduction

In the wake of the recent boom in the field of Artificial Intelligence, there has been an equally spectacular boom in apocalyptic predictions regarding AI and the faith of mankind. Extrapolating the accelerating progress of AI and our dependence on it, doomsayers worry that it is only a matter of time before we develop an Artificial General Intelligence, or Strong AI, which would be so powerful that it could cause terrible global suffering and possibly even the extinction of our species (Bostrom, 2014; Hawking et al, 2014; Tegmark, 2015; Musk, 2015; BBC, 2015). In fact, our situation is deemed so worrisome, that several new research centers have been created with the explicit aim of reducing the potential dangers of AI.[1]

Some authors have also turned the table on the ethical concerns regarding AI. Instead of merely considering the harm that an AI could bring upon humans,

[1] The Center for Human-Compatible AI, the Machine Intelligence Research Institute, OpenAI, the Future of Humanity Institute, and the Foundational Research Institute, to name just a few. Of course these institutes do not focus exclusively on the long-term existential risks posed by AI, but also on the abundant more concrete risks that current AI already poses.

they also consider the harm that could be brought upon an AI by humans (Metzinger, 2010; Bostrom, 2014; Mannino et al, 2015; Sotala and Gloor, 2017). The idea is that a truly intelligent AI would also develop consciousness, and with consciousness comes the capacity for emotions, agency, and all other aspects that we associate with subjects that deserve moral consideration (Dennett, 1993; Chalmers, 1996; Bostrom, 2014; Metzinger, 2010). For example, one could argue that the continued development of AI as systems that are entirely subjected to our every wish and command would amount to re-introducing slavery (Walker, 2006).

Ironically, the strong pessimism towards the future prevalent in both types of ethical concern mentioned above is founded on underlying assumptions that reveal a strong optimism towards the present: the assumption that the current rate of progress within AI is bound to continue unabated and the assumption that we have a clear understanding of certain deep philosophical issues. We set aside entirely whether the former assumption is justified, as that is something to be settled by a technical scientific discussion. Instead, the focus of this paper is on the latter more philosophical assumptions.

Concretely, the first type of ethical concern is based on the assumption that an AI could become superintelligent and the second type of ethical concern is based on the assumption that an AI could suffer. Both assumptions are highly controversial from a philosophical perspective. Firstly, it is not at all clear whether the very notion of superintelligence makes any sense, especially as it concerns non-human entities. Secondly, it is undoubtedly an understatement to say that we do not yet have a good understanding of how consciousness arises in human beings, let alone elsewhere.

The goal of this paper is not to call into question these assumptions directly. Instead, the aim is to show that if we actually take them seriously, then humanity has a moral obligation not to create a conscious AI. If this argument is successful, researchers in AI who are reluctant to accept its conclusion will be pressed with the challenge of either dropping one – or both – of these assumptions, or do some soul-searching and become advocates of a ban on the creation of a conscious AI.

In addition to said assumptions, the argument here developed also assumes a minimal utilitarian outlook. That is, it is assumed that utilities appropriately capture certain quantifiable and objective features of our moral framework, and that all else being equal, we have a moral duty to create more utility rather than less. What makes this outlook minimal, is that it leaves open entirely whether utilities capture *all* morally salient features.

2 Supersuffering

The first assumption we encountered above is that an AI could develop superintelligence, which is a level of intelligence that far exceeds our own human intelligence (Chalmers, 2010; Bostrom, 2014). In the words of Bostrom, "we mean an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" (Bostrom,

2006). Although it is quite straightforward to imagine an AI that outperforms human beings in computationally demanding tasks, it is much harder to conceive of an intellect that is much smarter than human beings in, say, social skills. Presumably we are talking about the skills required to socialize with human beings, after all.

More generally, given that the concept of intelligence was constructed to capture a property of human beings, it is not at all clear what it even means to surpass *human* intelligence. The way in which the idea of superintelligence is invoked implies that it isn't merely a matter of being able to think faster than a human or having a larger memory capacity, for if that were the case then humans would still be able to outsmart it by using their collective intelligence together with computers and other technology. On the contrary, the type of superintelligence produced by a singularity, or one that has the capacity to subdue and potentially destroy all of mankind, is such that it would be able to gain insights that are seemingly forever beyond our grasp. Here one is inclined to paraphrase Wittgenstein: "If an AI could speak, we could not understand it".[2]

Without a proper theory of intelligence that enables us to make sense of the idea that intelligence comes in degrees which extend far beyond the range of anything we find in humans, we ought to split up our single assumption into several separate assumptions that make explicit what the idea of a superintelligent AI requires.

Assumption 1 (Mental) *There exist quantifiable, mental, and human properties that an AI can have.*

Assumption 2 (Super) *If an AI can have a quantifiable, mental, and human property, then it can have this property to a degree which extends far beyond the human level.*

Assumption 3 (Intelligence) *Intelligence is one such property for which Assumption 1 holds.*

Taken together, these assumptions allow us to conclude that a future AI could be superintelligent.

The second assumption mentioned earlier is that an AI could suffer. Given that, like intelligence, suffering comes in degrees, we can rephrase this assumption as follows:

Assumption 4 (Suffering) *Suffering is one such property for which Assumption 1 holds.*

As with intelligence, we can apply Assumption 2 to conclude that a future AI could supersuffer, i.e., it could suffer to a degree that far exceeds any potential human suffering.

One might object that Assumption 2 is too strong, for the idea of superintelligence only requires an assumption of that form to hold for intelligence.

[2] The original mentions a lion, rather than an AI. (Wittgenstein, 1953, p.223)

Yet without a theory of intelligence that gives us this particular assumption, restricting Assumption 2 to intelligence would be gratuitous: we have no grounds whatsoever for stipulating that there is something peculiar about intelligence as compared to suffering so that it is the only candidate for amplification to a super-level.

Further, we find additional support for the possibility of a supersuffering AI from other sources. Sotala and Gloor (2017) offer a detailed analysis of the potential suffering that could be caused by an AI. While they focus mostly on human suffering, they also mention that "these [future technologies] may enable the creation of mind states that are worse than the current biopsychological limits." They provide interesting thought experiments to substantiate this claim. In a similar vein, Metzinger states that future AI's "might suffer emotionally in degrees of intensity or in qualitative ways completely alien to us that we, their creators, could not even imagine." (Metzinger, 2013, p. 6).

Such a supersuffering AI would amount to what can be called a *negative utility monster*: a being whose utility is so incredibly low that all of our efforts should go to increasing its utility, instead of wasting energy on increasing the comparatively negligible utilities that we human beings could obtain. The notion of a positive utility monster was posited by Nozick in order to highlight a counterintuitive consequence of utilitarianism (Nozick, 1974, p. 41):

> Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater sums of utility from any sacrifice of others than these others lose ... the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility.

One standard utilitarian reply is to object that such a monster is not conceivable, for no single entity could possibly have such large quantities of utility, be it negative or positive (Parfit, 1984). Note that our starting point, however, contains the observation that the recent success of AI has dramatically altered the type of entities that people claim they can conceive of. So if by now we can conceive of an AI as an intelligence monster, and we can conceive of an AI as having morally salient mental states such as suffering, then the mere claim that we cannot conceive of an AI as a negative utility monster does not carry much weight.

Other utilitarians are indeed prepared to bite the bullet and concede that such a monster would have to be the primary target of moral concern. For example, Singer says that "if we ever encountered Martians who could convince us that they had a vastly greater capacity for happiness than we do, then it could be a problem." (Singer, 2009).

So far we have focussed on the suffering of a single AI, but the problem of supersuffering becomes all the more pressing once we aggregate the suffering of multiple AI systems, over extended periods of time. Once we are able to create an AI – in the sense of a superintelligent and conscious AI as we have been considering – it is reasonable to assume that we go ahead and produce a large number of copies. From there it is only a small step to imagining horrible sce-

nario's in which there would be more artificial suffering than all human suffering, past and present, combined.

For example, say we are able to create a holographic AI that is the result of uploading a person's brain and running it as a hologram that looks like the person. Creative as human beings are, a cunning investor uses this technology to construct a profitable attraction: for a couple of dollars, visitors get to pull the switch on a holographic electric chair in which is seated a holographic copy of a convicted murderer whose original human version has long since been executed by an actual electric chair. The fact that the hologram experiences the exact same excruciating pain makes the attraction widely successful. Millions of visitors come to pull the switch, causing millions of holograms to suffer terribly. To top it all of, each visitor receives a keychain containing a copy of the hologram that is continuously, during every single second of the day, year after year, experiencing this execution. Unlike an actual human being, these holographic AI's do not have the benefit of death to put their suffering to an end. Obviously this scenario is extremely far-fetched, but given our earlier assumptions, it is certainly conceivable. This is confirmed by millions of viewers of the superb sci-fi television series Black Mirror, in which this very scenario is enacted (and others like it).

For another illustration, one could imagine that the experience of empathy is achieved in an AI by automatically replicating any suffering that it observes. A reason for programming the AI in this manner is that it might very well be a good way to ensure that an AI is highly sensitive to, and aware of, any form of human suffering – which it better be if we expect it to avoid treating humans as mere instruments for attaining its objectives. Now imagine that such an AI has access to all of recorded human history. In particular, it can immediately access all audio-visual material ever produced. Further, the AI is so fast and unbounded in resources that for every single decision it makes, it takes into account the total amount of evidence which is available to it. Say it makes a million decisions per second. This implies that during a single second, a single AI goes through the entire amount of suffering ever recorded a million times over. The fact that all of this happens within a single second should not be seen as a mitigating factor, for according to Bostrom and Yudowsky's plausible principle of the subjective rate of time, "In cases where the duration of an experience is of basic normative significance, it is the experience's subjective duration that counts." (2014, p. 326). Given the speed at which we can expect an AI to be operating, this principle in and of itself is already sufficient to guarantee that the experience of suffering for an AI can take on far more extreme forms than it can for human beings: a single experiment that goes astray for a few seconds could result in an AI suffering for many years.

One might counter that we can avoid such scenario's by implementing policies that forbid them. But such policies would be unable to prevent similar scenario's in which the suffering is unintended, and worse even, scenario's in which the suffering goes by entirely unnoticed. Once an AI has the capacity to suffer, then all it would take is some bug in the code for similar scenario's to unfold. For

example, imagine that there is some complicated version of the millennium bug, which is activated in billions of AI's at the same time and causes them to suffer to the astronomic extend portrayed above before we even know what is going on.

Metzinger also focusses on this issue, highlighting the "possibility that non-biological subjects of experience have already begun to suffer before we as their human creators have even become aware of this fact." (Metzinger, 2013, p. 6). He develops a theory that allows for the quantification of suffering, and posits that it is our duty to minimize the frequency of conscious experiences that involve suffering (Metzinger, 2017). As a consequence, he concludes that we should ban the development of an AI, in the strong sense of AI as we are using it, stating the following principle (Metzinger, 2013, p. 3):

Principle 1 Negative Synthetic Phenomenology (NSP) *We should not deliberately create or even risk the emergence of conscious suffering in artificial or postbiotic agents, unless we have good reasons to do so.*

Mannino et. al. reach a similar conclusion in their overview of the moral risks posed by the development of AI, stating that "the (unexpected) creation of sentient artificial life should be avoided or delayed wherever possible, as the AI's in question could – once created – be rapidly duplicated on a vast scale." (Mannino et al, 2015, p. 10).

Given the assumptions made at the outset, and the severity of the sketched scenario's, the only way to avoid accepting these negative verdicts is to follow through on Metzinger's hint and offer good reasons as to why the possibility of supersuffering is an acceptable price to pay. Three straightforward suggestions present themselves as plausible candidates:

1. The attempt at creating an AI is not at all special in this regard, since all other acts that we perform as humanity today are also possible causes of extreme suffering in the future, and nevertheless we find this perfectly acceptable.
2. The negative scenario of supersuffering is compensated by a positive scenario of an AI experiencing superpleasure.
3. The expected benefits for mankind that come from creating an AI outweigh the possibility of supersuffering.

In the remainder of this paper the aim is to show why all three suggestions fail.

3 The Unique Responsibility of Creating an AI

In order to show how the consequences of creating an AI are unlike the consequences of other acts that we collectively engage in, we focus on a particularly strong form of responsibility that leaves no room for excuses. If an agent causes some terrible outcome O, then she might avoid being held responsible for causing O if she had good reason to believe that the action she performed was far less

likely to cause O than any alternative available to her. For example, if a doctor injects medication into a patient in order to cure a life-threatening disease, and the patient dies due to an allergic reaction which the doctor could not possibly have foreseen, then the doctor would not be responsible for the patient's death under the strong form of responsibility we have in mind.

Definition 1 (Definite Responsibility). *If our actions cause an outcome O, then we are* definitely responsible *for O if we knew that our actions could possibly cause O and we knew that there were alternative actions that would certainly not have caused O.*

We already concluded that humanity might cause future AI's to experience supersuffering. In order to invoke the above definition, we need to add the following trivial counterpart.

Premise 1 *If all of humanity does not attempt to create an AI, then the set of our acts will certainly not cause an AI to ever experience supersuffering.*

This leads us to conclude the following:

Conclusion 1 *If (current) humanity creates AI systems, then we will be definitely responsible for all supersuffering (and superpleasure) they might endure in the future.*

On the short term, and when considering a single agent, there are many outcomes for which one is definitely responsible. This no longer holds if we consider all of humanity and extend our horizon into the far future: given our limited knowledge of the world, and the almost infinite complexity of the causal chain that results from our actions, we are ignorant with respect to the long-term consequences of our actions on the well-being of humanity.

Premise 2 *There exists a time t such that for any set of acts A that we perform today, to the best of our knowledge, it is possible that A will cause extreme human suffering after t.*

At first glance there appear to be many actions that defy the above premise. For example, think of our efforts to cure cancer. Either these are successful, in which case they would prevent a great deal of suffering, or they would be unsuccessful and not have any impact at all. But this analysis only considers the most likely outcomes that each alternative would have. Although unlikely, it is definitely possible that our efforts to cure cancer result in the creation of a deadly and contagious virus, which causes many more deaths than the disease which it was supposed to cure. Or it is possible that by curing cancer, the next would-be genocidal dictator is kept alive, and therefore able to live out his evil intentions.

There is one important type of actions that does form a convincing exception to this premise: if humanity stops having children altogether, we can be certain that we will not cause any human suffering in the long run. Therefore one might object that this line of reasoning opens the door to the extremely controversial

position of anti-natalism, championed by Benatar (2006), according to which it is immoral to have children. Indeed, we are definitely responsible for any suffering that our children may experience, because our children would not exist without us performing certain actions. We come back to this objection later.

We can now apply Definition 1 to reach the following conclusion:

Conclusion 2 *There exists a time t such that (current) humanity will not be definitely responsible for any human suffering (or pleasure) that occurs after t.*

This conclusion rules out the first suggested candidate as a good reason for risking supersuffering: the reason why we find it acceptable that our acts might have extremely negative consequences for humanity in the future, is that *this holds just as well for any alternative acts that we might perform.* All we can do is focus on outcomes in the near-future, and hope for the best in the long-term. The distinguishing feature of our attempt at creating an AI is that this is no longer true, for *there is an obvious alternative act* that will certainly not cause supersuffering, namely to stop doing any research on AI.

4 Moral Asymmetry

At this point we can draw the following worrisome conclusion.

Conclusion 3 *It is possible that by creating an AI, we will be definitely responsible for the greatest suffering that our world has ever known.*

Still, an optimist might argue, completely analogous to this depressing conclusion, we could also be responsible for the greatest pleasure that our world has ever known. Hence the route for the second suggestion to defend our attempt at creating an AI is still open.

However, there is a strong intuition that is so well-embedded in our everyday life that only an extreme utilitarian would object to it: it is more important to avoid suffering than it is to create pleasure. Moore (1903) was the first to express this intuition, but Popper was its most famous defender (Popper, 1945):

> We should realize that from a moral point of view suffering and happiness must not be treated as symmetrical; that is to say the promotion of happiness is in any case much less urgent than the rendering of help to those who suffer, and the attempt to prevent suffering.

This idea forms the basis of "moderate negative utilitarianism", which considers it our primary duty to avoid suffering (Parfit, 1997; Mayerfeld, 1999; Metzinger, 2013; Chauvier, 2014). The asymmetry between pleasure and pain that lies at its core is evident in the medical principle "first do no harm", and is confirmed by the moral risk-aversion that is widespread in our behaviour. [3]

For example, assume you may press a button such that with probability 0.5 a random person's leg will be broken, and with probability 0.5 someone's broken

[3] See the papers cited above for many more interesting examples.

leg will be healed, and neither person has any say in the matter. Or imagine that if you press the button, a random person will be hit in the face, but offered a massage afterwards. It goes without saying that it is immoral to press the button.

Further, this asymmetry increases as the intensity of the suffering and pleasure increases. For example, if someone insults you but then offers a compliment, you probably will not have hard feelings towards that person. But if they torture you, it is hard to imagine what form of pleasure they could offer you to avoid feeling terribly wronged by that person. In fact, some even go so far as to state that certain amounts of suffering can not be compensated by any amount of pleasure at all, a position Hurka describes as the limit asymmetry thesis: "There is some intensity n such that a pain of intensity n is more evil than any pleasure could be good." (Hurka, 2010, p. 205).

In light of all this, the following moral principle is endorsed by a broad range of ethical positions and has a prima facie intuitive appeal:

Principle 2 (Moral asymmetry) *All else being equal, the moral blameworthiness for being responsible for a degree of suffering X is greater than the moral praiseworthiness for being responsible for a degree of pleasure X. Further, the difference between the degree of blame and praise strictly increases with X.*

Nevertheless, as said, a strict utilitarian could insist on the symmetry between pleasure and suffering, and hence reject this principle. In that case, the possibility of a supersuffering AI could be compensated, on the condition that the probability of superpleasure is significantly greater than that of supersuffering.

Setting this caveat aside, we conclude that our expected blameworthiness when continuing the development of AI is higher than when we stop all research on AI, and hence the second candidate suggested as a good reason for risking supersuffering is ruled out as well.

5 The Ethical Priority of Artificial Suffering

The third suggestion that might offer a good reason for going ahead and risk the prospect of AI systems supersuffering, is that this would be an acceptable price to pay given the expected benefits for mankind that would follow the invention of truly intelligent AI. That we consider this candidate suggestion last is due to the simple reason that the discussion of the previous two suggestions already puts considerable pressure on this idea.

First, whatever increase in human pleasure the invention of an AI might cause, we would not be definitely responsible for any pleasure that lies beyond the time horizon t mentioned in Conclusion 2. In fact, if the doomsayers mentioned in the beginning are taken seriously, we have every reason to believe that the expected benefits for mankind in the long run are drastically negative. Even without invoking these worst-case scenario's, given the myriad degrees of freedom that the organization of human society possesses, there is no reason to assume that AI is in any way necessary for human beings to flourish in the long run. For

all we know, there exist other forms of actions not involving the creation of AI that would cause an even higher increase in human pleasure after time t than an AI could ever produce.

Second, we can combine the asymmetry between suffering and pleasure captured by Principle 2 with the astronomical difference in orders of magnitude between the amount of suffering depicted in the scenario's from Section 2 and any feasible amount of human pleasure that could occur before time t mentioned above to see that the benefits for mankind before time t do not even come close to compensating for a supersuffering AI. That is, they do not come close if we accept the following plausible principle:

Principle 3 (Non-discrimination) *When evaluating the overall expected benefits of creating an AI, we ought not discriminate between the suffering/pleasure of an AI and a human being.*

When reformulated in terms of different groups of human beings, this principle is a bedrock of any modern moral system, and hence it hard to see how it could fail to apply when we extend it to other conscious beings that have an even stronger capacity for suffering and pleasure than humans do. Singer puts it thus (Singer, 2011, p. 50):

> If a being suffers, there can be no moral justification for refusing to take that suffering into consideration. No matter what the nature of the being, the principle of equality requires that the suffering be counted equally with the like suffering – in so far as rough comparisons can be made – of any other being.

In sum, the combination of these two arguments blocks the last suggested candidate:

Conclusion 4 *From an ethical point of view, the possibility for an AI to experience supersuffering takes precedence over the expected benefits that an AI will produce for mankind.*

5.1 Anti-natalism

As promised, we briefly return to the accusation that the analysis here developed supports anti-natalism. This charge loses its sting if we examine the underlying motivation for this position. We motivated Principle 2 by reference to moderate negative utilitarianism. By using the label "moderate", its proponents wish to distance themselves from "negative utilitarianism", which embraces the far stronger claim that even the slightest amount of suffering can never be compensated by any amount of pleasure whatsoever. Given that every human being will experience some amount of suffering throughout its life, this claim implies that it is immoral to bring children into existence, no matter what the circumstances.

In contrast, a moderate negative utilitarian can perfectly well defend having children, on grounds of the fact that most people end up leading lives which have

an acceptable amount of suffering compared to the amount of pleasure. In other words, most people end up leading lives worth living. So the act of creating an AI is not similar to the act of having a child precisely because supersuffering is so extreme that it can not be compensated by superpleasure, as outlined in Section 4. Only if it were very likely that one's child would experience constant suffering would it follow that it is better not to have a child, a conclusion which most people would fully endorse.[4]

In addition, there is also a further obvious feature that distinguishes creating AI systems from creating future human beings, namely the fact that many people would find the prospect of mankind going extinct quite depressing, whereas few would mourn the non-existence of sentient AI's.

6 Conclusion

In Section 2 we examined the basis for two popular and controversial assumptions regarding Artificial Intelligence, and argued that accepting these assumptions leads to the moral principle that we should not create a conscious AI, unless we can offer good reasons to do so. In the subsequent sections we rejected three natural candidates for such reasons. Therefore, if we take seriously our two initial assumptions, we are forced to accept the following conclusion:

Conclusion 5 *Humanity should not attempt to create a conscious AI.*

Given the gravity of this conclusion, it is incumbent upon each and every AI researcher to closely inspect said assumptions, and make a choice: either refrain from endorsing both of them and explain why, or advocate for a ban on the creation of a conscious AI.

Acknowledgements

The author gratefully acknowledges financial support from the ERC-2013-CoG project REINS, nr. 616512

[4] Metzinger also makes this point, and adds that anti-natalism regarding artificial life is far more plausible than its biological counterpart (Metzinger, 2013, 2017). To avoid unnecessary complication, we make clear that we need not get into the issue of abortion, but are talking simply about preventing the act of human fertilization in the first place.

Bibliography

BBC (2015) Stephen hawking warns artificial intelligence could end mankind. URL http://www.bbc.com/news/technology-30290540

Benatar D (2006) Better Never to Have Been. Oxford University Press

Bostrom N (2006) How long before superintelligence? Linguistic and Philosophical Investigations 5(1):11–30

Bostrom N (2014) Superintelligence: Paths, Dangers, Strategies. Oxford University Press

Bostrom N, Yudowsky E (2014) The Cambridge Handbook of Artificial Intelligence, Cambridge University Press, chap The Ethics of Artificial Intelligence

Chalmers D (2010) The singularity: A philosophical analysis. Journal of Consciousness Studies 17(9-10):7–65

Chalmers DJ (1996) The Conscious Mind: In Search of a Fundamental Theory. Oxford University Press

Chauvier S (2014) A challenge for moral rationalism: why is our common sense morality asymmetric? In: Dutant J, Fassio D, Meyan A (eds) Liber Amicorum Pascal Engel, University of Geneva, pp 892–906

Dennett DC (1993) Consciousness Explained. Penguin UK

Hawking S, Russell S, Tegmark M, Wilczek F (2014) URL http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence-but-are-we-taking-9313474.html

Hurka T (2010) Asymmetries in value. Nous 44(2):199–223

Mannino A, Althaus D, Erhardt J, Gloor L, Hutter A, Metzinger T (2015) Artificial intelligence: Opportunities and risks. Policy Paper by the Effective Altruism Foundation 2:1–16

Mayerfeld J (1999) Suffering and Moral Responsibility. Oxford University Press

Metzinger T (2010) The Ego Tunnel: The Science of the Mind and the Myth of the Self. New York: Basic Books

Metzinger T (2013) Two principles for robot ethics. In: Hilgendorf E, Günther JP (eds) Robotik und Gesetzgebung, Baden-Baden: Nomos, pp 263–302

Metzinger T (2017) Suffering. In: Almqvist K, Haag A (eds) The Return of Consciousness, Axess Publishing

Moore G (1903) Principia Ethica. Cambridge University Press

Musk E (2015) URL https://www.theguardian.com/technology/2015/jan/16/elon-musk-donates-10m-to-artificial-intelligence-research

Nozick R (1974) Anarchy, State, and Utopia. Basic Books

Parfit D (1984) Reasons and Persons. Oxford University Press

Parfit D (1997) Equality and priority. Ratio 10(3):202–221

Popper K (1945) The Open Society and Its Enemies, Vol. I. Routledge

Singer P (2009) Back talk: Peter singer. URL https://www.thenation.com/article/back-talk-peter-singer/

Singer P (2011) Practical Ethics. Cambridge University Press

Sotala K, Gloor L (2017) Superintelligence as a cause or cure for risks of astronomical suffering. Informatica 41:389–400

Tegmark M (2015) An open letter: Research priorities for robust and beneficial artificial intelligence. URL https://futureoflife.org/ai-open-letter

Walker M (2006) A moral paradox in the creation of artificial intelligence: Mary poppins 3000s of the world unite! AAAI Workshop: Human Implications of Human-Robot Interaction

Wittgenstein L (1953) Philosophical Investigations