Journal of EVOLUTION & Technology

A peer-reviewed electronic journal published by the Institute for Ethics and Emerging Technologies

ISSN 1541-0099

29(1) – October 2019

Do No Harm Policy for Minds in Other Substrates

Soenke Ziesche

soenke.ziesche@gmail.com

Roman V. Yampolskiy
University of Louisville

roman.yampolskiy@louisville.edu

Journal of Evolution and Technology - Vol. 29 Issue 2 – October 2019 – pgs 1–11

Abstract

Various authors have argued that in the future not only will it be technically feasible for human minds to be transferred to other substrates, but this will become, for most humans, the preferred option over the current biological limitations. It has even been claimed that such a scenario is inevitable in order to solve the challenging, but imperative, multi-agent value alignment problem. In all these considerations, it has been overlooked that, in order to create a suitable environment for a particular mind – for example, a personal universe in a computational substrate – numerous other potentially sentient beings will have to be created. These range from non-player characters to subroutines. This article analyzes the additional suffering and mind crimes that these scenarios might entail. We offer a partial solution to reduce the suffering by imposing on the transferred mind the perception of indicators to measure potential suffering in non-player characters. This approach can be seen as implementing literal empathy through enhanced cognition.

Introduction

Due to recent technological progress, it appears to have become more realistic to enhance human minds or even transfer them to other substrates. In this introduction, we set out four assumptions, followed, in the next section, by formulating a problem to which they lead. In summary, we argue that enhancement and substrate-transfer scenarios are 1) desirable, 2) may become feasible, 3) could even be inevitable in order to tackle the multi-agent value merger toward AI safety, but 4) may affect other sentient minds.

1) Desirability: The transhumanist movement has for some time advocated the enhancement of human minds (e.g. More 2013). Bostrom illustrates the desirability of enhanced human capacities by describing potential enhancements related to health span, cognition, and emotions (Bostrom 2008). The potential scenario when the quality of virtual worlds has

reached a level where human minds prefer them to the physical world has been called by Faggella (2018b) "Programmatically Generated Everything."

2) Feasibility: Two main categories are distinguished here (see, for example, our discussion in Yampolskiy and Ziesche 2018): Virtual worlds comprise virtual and augmented reality through ever-improving devices that are experienced by a biological human mind. Uploads refer to the potential transfer of human minds to other physical substrates, for example a computer. While virtual worlds have been implemented already with progressing quality (e.g. Faggella 2018a), the feasibility of uploads has also been suggested, for example by Sandberg's and Bostrom's roadmap for whole brain emulation (Sandberg and Bostrom 2008) and some others (e.g. Koene 2012; Tegmark 2017).

3) Inevitability: AI safety is of paramount importance and requires undertaking various challenges, of which the multi-agent value merger within the multi-agent value alignment problem is one of the hardest. As a solution, Yampolskiy (2019) proposes Individual Simulated Universes (ISUs), which are personalized simulations created by superintelligent AIs for all human minds. Yampolskiy argues that this approach would have the additional benefit of providing unprecedented potentials as well as more and lasting happiness to the human minds experiencing ISUs. This affirms the assumed desirability of such an endeavor.

4) Involvement of other sentient minds: Various authors argue that already now, or in the future, sentient digital beings or minds may exist and they may, for example, constitute subroutines as well as non-player characters (NPCs) in video games, simulations, or other computational substrates (Bostrom et al. 2018; Bostrom 2018; Tomasik 2019a; Tomasik 2019b; Ziesche and Yampolskiy 2019). This implies that computational substrates for the enhancement and transfer of human minds will also contain other sentient beings, since NPCs and subroutines are essential components of them.

Problem formulation

It has been argued that sentient digital minds have a moral status because of their feature of being sentient (Bostrom 2018; Bostrom, Dafoe, and Flynn 2018; Tomasik 2019a; Tomasik 2019b; Ziesche and Yampolskiy 2019). However, in the discussion about enhancement and transfer of human minds to other substrates the focus is usually on the advantages and opportunities for human minds, while any potential suffering experienced by the sentient beings inherent to these substrates has mostly been neglected.

Tomasik recently formulated the problem as follows:

> Imagine a posthuman paradise in which advanced human-like beings are simulated in blissful utopian worlds, never experiencing (access-conscious) suffering. Their minds might nonetheless contain suffering subroutines, such as neural signals that fail to win control of action, or signals within cognitive modules that are inherently inaccessible to explicit report. In addition, the machines running such simulations might themselves contain suffering subroutines, such as in their operating systems. (Tomasik 2019a)

Although this is speculative, in such a scenario the overall suffering per computational substrate might outweigh the bliss of the transferred human mind, which defeats the original purpose. This would actually be one example to support concerns that technical developments may increase risks of astronomical suffering (e.g. Sotala and Gloor 2017).

Before moving on, we will turn to potential objections regarding the assumptions and the problem.

Could it be possible that neither subroutines nor NPCs are sentient? Yes, this could be possible since sentient digital minds are speculation. Simple subroutines or NPCs, which consist of some if-statements only, are probably non-sentient, hence a comment by Bostrom, Dafoe, and Flynn:

> Policymakers are unaccustomed to taking into account the welfare of digital beings. The suggestion that they might acquire a moral obligation to do so might appear to some contemporaries as silly, just as laws prohibiting cruel forms of recreational animal abuse once appeared silly to many people. (Bostrom, Dafoe, and Flynn 2018, 16)

Given human beings' track record of causing immense suffering thanks to recognizing ethical issues too late, and in order not to repeat such mistakes, we should step cautiously here. The potential suffering of sentient digital minds in computational substrates created for the enhancement and transfer of human minds should be given serious consideration and be addressed in a timely way (Bostrom, Dafoe, and Flynn 2018; Ziesche and Yampolskiy 2019). We assume, therefore, there might be sentient subroutines and NPCs in computational substrates, and this subset of subroutines and NPCs provides the focus for what follows.

Could it be possible to create such computational substrates without subroutines? The answer has been provided by Tomasik: "Eliminating suffering on the part of simple computational processes seems impossible, unless you dispense with computation altogether" (Tomasik 2019a).

Could it be possible to create such computational substrates without NPCs? In theory, computational substrates for the enhancement and transfer of human minds devoid of any NPCs are possible, but it then becomes very questionable whether our *desirablility* assumption is fulfilled. Yampolskiy (2019) has proposed ISUs in order for human minds to be happy, and, perhaps with a very few exceptions, it is hard to imagine human minds being enduringly happy without any social interaction with other minds.

Therefore, we face a challenge: given the desirability, feasibility and inevitability of ISUs, how can the suffering of other sentient beings be avoided, or at least reduced, in computational substrates for the enhancement and transfer of human minds?

Typology of relevant minds

The space of all minds has been described as vast (e.g. Sloman 1984; Yudkowsky 2008; Yampolskiy 2015). In order to tackle the problem as we've defined it, we first present a typology to establish which subset of this vast space might comprise the relevant computational substrates. As indicated, we distinguish two main categories of sentient digital minds: NPCs and subroutines.

NPCs

The term "non-player character" originated in the realm of gaming and has been defined as any character that the player does not control. In recent times, the complexity of NPCs has evolved significantly, and the concept has also been transferred to virtual worlds and simulations. Tomasik has broached whether NPCs matter morally (Tomasik 2019b), while Warpefelt and Verhagen have presented a suggestive typology, based on the video game domain, with the following roles for NPCs:

> Buy, sell and make stuff, provide services, provide combat challenges, provide mechanical challenges, provide loot, give or advance quests, provide narrative

> exposition, assist the player, act as an ally in combat, accompany the player, and make the place look busy. (Warpefelt and Verhagen 2015, 7–8)

Such existing typologies are, however, much too narrow, as well as too anthropomorphic to classify the NPCs likely to be found in upcoming environments for enhancement and transfer of human minds. Detailed typologies are not possible at this point, since future NPCs may be unimaginably alien, given that in future virtual worlds and ISUs basically anything might be possible (e.g. Loosemore 2014; Faggella 2018a; Yampolskiy 2019).

What matters here instead is the question: What might cause NPCs to *suffer* in virtual worlds and ISUs? Three categories can be distinguished:

- The enhanced/transferred human mind intentionally causes NPCs to suffer.
- The enhanced/transferred human mind unintentionally causes NPCs to suffer.
- NPCs suffer, but this is not caused by actions of the enhanced/transferred human mind.

The first category resembles the concept of mind crime, introduced by Bostrom with AIs as the perpetrators (Bostrom 2014). In this case, the enhanced/transferred human mind knows about the consequences of her or his activity but experiences sadistic pleasure or has other objectionable motivations.

For the second category, suffering might be caused by the alien features of the NPC, as a result of which the enhanced/transferred human mind is not aware that he or she is causing suffering. As a result of the activities of the enhanced/transferred human mind, the NPC might undergo aversive sensory experiences that the enhanced/transferred human mind cannot imagine.

The third category comprises potential suffering by NPCs when not interacting with the enhanced/transferred human mind. For example, the NPC might be suffering from boredom because of a different subjective rate of time, which could be an "exotic property" of NPCs (Bostrom and Yudkowsky 2011). NPCs might, moreover, harm each other, thereby causing suffering. In addition, there might be as many more potential ways of suffering as there are possibly unknown unknowns regarding aversive sensory experiences of digital minds.

Another helpful distinction would be between friendly or neutral NPCs and hostile NPCs, since the intentional causation of suffering toward hostile NPCs by the enhanced/transferred human mind might be considered self-defense.

Subroutines

Given the lack of evidence, it is challenging to develop a typology of subroutines that relates to suffering in computational substrates for the enhancement and transfer of human minds. Here we can distinguish whether the subroutines are executed within the mind of the transferred human or in other parts of the computational substrate. The latter require further specification as those subroutines that do not constitute NPCs (since NPCs have already been discussed). Again, based on the possibility of very alien NPCs, this distinction is not simple: that is, there might not be a clear-cut line as to what features are required for subroutines to count as NPCs. Nevertheless, for our current purposes this is not a problem since we aim to explore the prevention of suffering for both NPCs and subroutines. Note, however, that for non-NPC subroutines there appears to be no scenario in which an enhanced/transferred human mind could intentionally cause or prevent suffering, regardless whether the subroutines are within or outside her/his mind.

Partial policy solution

In a recent paper, Bostrom and his collaborators formulated the desideratum "that maltreatment of sentient digital minds is avoided or minimized" (Bostrom, Dafoe, and Flynn 2018, 18), and elsewhere Bostrom has encouraged addressing this issue early "while the artificial agents we are able to create are still primitive" (Bostrom 2018, 2). As a follow-up, we recently termed this field of research "AI Welfare Science" (Ziesche and Yampolskiy 2019). The aim here is to reduce or prevent the suffering, as well as the unwanted deletion, of digital sentient minds. At the same time, we offered recommendations for AI welfare policies. Sotala and Gloor have also presented recommendations on this issue (Sotala and Gloor 2017, 10).

Since there is no evidence that digital minds are incapable of sentience or immune to suffering, and since AI Welfare Science, which is in its very early stages, has not yet developed methods to abolish suffering of digital minds, policies are required to prohibit an enhanced/transferred human mind from causing suffering.

Owing to the alienness of the new environment, the enhanced or transferred human mind is likely to face challenges in its efforts to identify the suffering of the NPCs with which she/he is interacting. Suffering might be observed through either physiological/functional or behavioral indicators (see our discussion in Ziesche and Yampolskiy 2019).

Behavioral indicators, which comprise self-reporting, have the disadvantage, in both the real world and a computational substrate, that they can be deliberately faked, which could include the possibility of non-sentient NPCs presenting as sentient and suffering.

Physiological/functional indicators, by contrast, are more impartial, and hence more suitable for use in real life for objective pain assessment in humans and non-human animals (e.g. Cowen et al. 2015). In their 2015 paper, Cowen's team studied markers to measure pain, such as changes in the autonomic nervous system, biopotentials, neuroimaging, biological (bio-) markers, and composite algorithms. Although the identification of parameters that correlate with pain intensity is challenging, progress has been made, and, for example, the nociceptive flexion reflex turned out to be a reliable and objective tool for measurement of pain (Skljarevski and Ramadan 2002).

Transferred to computational substrates, this issue should be more tractable as, unlike in the real world, everything is measurable precisely as well as constantly. If suffering, such as pain, can also affect digital minds, then there must be quantitative indicators for this, which are called here "computational." This leads to our main proposal for a *do no harm policy for minds in other substrates*:

> For the development of computational substrates that have the purpose of accommodating human minds, it is mandatory that the transferred human mind in such a substrate be equipped with sensory perception, through which she/he perceives computational indicators of suffering of the NPCs with whom she/he interacts.
>
> If the stimuli of these indicators reach the threshold of suffering, the human mind ought to stop any activities that cause the suffering.

Since exploring qualia is a difficult problem (Chalmers 1995), we emphasize quantitative and objective physiological/computational indicators. If there was an option to let the transferred human mind compulsorily perceive directly any unpleasant qualia he/she might causing to local NPCs, this would be an even stronger tool to prevent the transferred human from doing harm, but this is too speculative at this stage.

Either way, through this policy the term *empathy* is taken literally, since the transferred human mind would perceive precise indicators of the effects of his/her actions toward NPCs. Moreover, the policy can be seen as an attempt toward mind crime prevention. This approach is also in line with our previous AI Welfare Science recommendations, which encourage developing methods to measure the suffering of sentient digital minds through physiological/functional or behavioral parameters (Ziesche and Yampolskiy 2019).

In the following specifications, we elaborate opportunities and gaps, as well as the bigger picture and future work related to the proposed do no harm policy.

Specifications

The following points further elaborate the policy:

- Through this policy, the first two categories of suffering in virtual worlds are covered. For *intentional maltreatment* of an NPC, passing the threshold of suffering indicators will be expected by the enhanced/transferred human mind, and will ideally be avoided. In cases of *unintentional maltreatment*, the enhanced/transferred human mind will need to learn about, and understand, the aversive sensory experiences of the affected NPC.

- The alternative option would be to simply prohibit enhanced/transferred human minds from harming other NPCs – similar to much legislation in the real world. However, this may not be workable without the sensory perception of suffering indicators, since the enhanced/transferred human mind, even if she/he has the best intentions, likely does not know sufficiently when and how NPCs suffer. The barrier to understanding is the likely alienness of an NPC's mind. The sensor should, therefore, be embraced as an opportunity provided by the potential of the computational substrate**.**

- It could be argued that the sensor triggers only once the pain of the NPC has commenced, and so the policy does not thoroughly prevent suffering. However, similarly to the development of infants, the enhanced/transferred human mind will gain experience literally through machine learning over time. Ideally, it will learn to prevent an NPC's suffering before it occurs.

- If the enhanced/transferred human mind does not obey the rules, he/she may be punished. If passing the threshold of the suffering indicators does not trigger the enhanced/transferred human mind to stop causing pain to an NPC, an option could be to stop the activities that cause suffering automatically – taking control from the enhanced/transferred human mind that has breached the do no harm policy. In this case, the degree of freedom within an ISU or other computational substrate – which is anyway an illusion – would be restricted.

Opportunities

The proposed do no harm policy incorporates various opportunities relating especially to the important topic of empathy:

- Empathy is a notable human ability to detect manifestations of distress in other humans (and to some extent in non-human animals). However, the accuracy of human empathy varies, particularly when it comes to animals. The do no harm policy harnesses the potential of computational substrates to allow for enhanced cognition, thus optimizing precise empathy to the extent of the literal meaning of the word. An

additional positive effect of this approach is that it overcomes the anthropomorphic bias of empathy, a step that is necessary for interaction with alien minds.

- It appears that the function of empathy for humans is to prompt action to reduce the suffering of the other mind, provided this is within his/her capacity. However, it has been claimed that this step is, in real life, not taken inevitably (e.g. Brooks 2011). The do no harm policy addresses this issue, insofar as it is more likely that the enhanced/transferred human mind will attempt to end the NPC's pain if she/he is obliged to perceive the extent of the pain. (Enhanced/transferred human minds with sadomasochistic tendencies might constitute a troubling exception.) This kind of artificially enhanced empathy can be seen as a building block for artificial conscience (e.g. Pitrat 2013).

- Due to the almost unlimited options in computational substrates such as ISUs, these could also be programmed so that an enhanced/transferred human mind perceives not only the aversive sensory experiences of the NPCs that it interacts with, but also NPCs' pleasant experiences. Thus, artificially enhanced empathy would cover the whole range of experiences of NPCs. However, that might be up to the enhanced/transferred human mind to decide. In order to create conditions that prevent suffering, it suffices if the enhanced/transferred human mind perceives the aversive experiences.

- Elsewhere, we treat the unwanted deletion of sentient digital minds as a distinct topic within AI Welfare Science (Ziesche and Yampolskiy 2019). Regarding NPCs and subroutines in virtual worlds and ISUs, this issue is again more straightforward for NPCs than for subroutines. The unwanted deletion of NPCs by an enhanced or transferred human mind should be forbidden apart from exceptional cases of self-defense. By contrast, the deletion of subroutines might be unavoidable, and it might often be neither controlled nor noticed by the enhanced/transferred human mind.

Gaps

The policy proposal is partial and at an early stage since, for example, it does not address the following cases:

- As mentioned above, a special case would be hostile or evil NPCs. Options would be to exclude hostile NPCs entirely (by a separate policy) or to accept causing suffering to them as self-defense.

- A way must be found to prevent NPCs from harming each other, as this would increase the overall suffering in the computational substrate.

- In this regard, two other undesirable activities of potentially violent or sadistic enhanced/transferred human minds have to be considered. Such a mind might enjoy observing NPCs harming each other – similar to human enjoyment of a cock fight – or might instruct NPCs to maltreat other NPCs. In either case, the policy would be bypassed and the sensor would be not triggered since the human mind would not be causing the suffering directly.

- Furthermore, NPCs might still suffer for reasons not linked to actions by an enhanced/transferred human mind or by other NPCs, for example because of boredom or alien aversive perceptions.

- It could be an option that the sensor detects the aversive experiences of these NPCs and the enhanced/transferred human mind thus be encouraged to help them, just as altruistic humans help others without having caused the misery. However, this 1) might be challenging for alien types of suffering and 2) might to some extent defeat the purpose of the virtual world or ISU for the enhanced/transferred human mind to enjoy complete freedom.

- Lastly, the do no harm policy does not cover the even-more-elusive potential suffering of subroutines.

Bigger picture and future work

Not only because of the gaps we have outlined, our policy proposal should be seen as only a beginning, to be embedded in a bigger picture and with various directions for future work:

- Yampolskiy (2019) has described a superintelligent AI exercising control of ISUs, which would likely also apply to other computational substrates sophisticated enough to simulate human minds. Such an AI would also be essential for the implementation of the policy proposed here, and therefore ensuring the safety and friendliness of the AI is crucial as well as challenging (Yampolskiy 2018; Yudkowsky 2008).

- In an earlier paper (Ziesche and Yampolskiy 2019), we proposed that the overarching goal should be suffering-abolitionism as elaborated by Pearce (2007), yet transferred to digital environments and ISUs in particular, which Pearce did not incorporate. Since suffering-abolitionism has not yet succeeded, and since the prevention of suffering has a moral urgency, we have proposed the policy sketched above.

- In the real world, research is being conducted toward crime prediction through big data such as mobile phone data (e.g. Bogomolov et al. 2014). This kind of "mind crime prediction" could also be applied to a computational substrate where the available data are much more abundant and precise, since the potential culprits are permanently monitored and recorded like every computation in the substrate.

- Artificially enhanced empathy could provide further positive effects. For example, this scenario could be envisaged as a training or cure for transferred human psychopaths or sociopaths.

- An interesting, but complex, challenge would be to compare the suffering caused by a human mind in the real world, for example by abjuring veganism, by killing insects, and so on, compared to the suffering caused by the same human mind after being enhanced or transferred to another computational substrate. Ideally, the latter would be less.

In summary, we have offered a partial solution to the problem of reducing or avoiding potential suffering of NPCs in computational substrates for the enhancement and transfer of human minds. Beyond this, we have sought to identify some neglected and remaining issues. The innovative concept and our demand for artificially enhanced empathy provide a contribution to urgently required AI policies and to AI safety.

Acknowledgment

Soenke Ziesche wishes to acknowledge the support of the Faculty of Engineering, Science and Technology within the Maldives National University, where he was employed while researching and drafting this paper.

References

Bogomolov, A., B. Lepri, J. Staiano, N. Oliver, F. Pianesi, and A. Pentland. 2014. Once upon a crime: Towards crime prediction from demographics and mobile data. Orig. pub. in *Proceedings of the 16th International Conference on Multimodal Interaction*, 427–34. ACM. https://arxiv.org/pdf/1409.2983.pdf (accessed October 7, 2019).

Bostrom, N. 2008. Why I want to be a posthuman when I grow up. In *Medical enhancement and posthumanity*, ed. Bert Gordijin and Ruth Chadwick, 107–36. Dordrecht: Springer. Available online https://nickbostrom.com/posthuman.pdf (accessed October 7, 2019).

Bostrom, N. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Bostrom, N. 2018. The interests of digital minds. Draft 1.0. https://nickbostrom.com/papers/interests-of-digital-minds.pdf (accessed October 7, 2019).

Bostrom, N., A. Dafoe, and C. Flynn. 2018. Public policy and superintelligent AI: A vector field approach. Version 4.3. https://pdfs.semanticscholar.org/9601/74bf6c840bc036ca7c621e9cda20634a51ff.pdf (accessed October 7, 2019).

Bostrom, N., and E. Yudkowsky. 2014. The ethics of artificial intelligence. In *The Cambridge handbook of artificial intelligence*, ed. K. Frankish and W. M. Ramsey, 316–34. Cambridge: Cambridge University Press.
Draft available online:
https://nickbostrom.com/ethics/artificial-intelligence.pdf (accessed October 7, 2019).

Brooks, D. 2011. The limits of empathy. *New York Times*. September 29. https://www.nytimes.com/2011/09/30/opinion/brooks-the-limits-of-empathy.html?scp=1&sq="the limits of empathy"&st=cse (accessed October 7, 2019).

Chalmers, D. J. 1995. Facing up to the problem of consciousness. *Journal of consciousness studies* 2: 200–19. http://cogprints.org/316/1/consciousness.html (accessed October 7, 2019).

Cowen, R., M. K. Stasiowska, H. Laycock, and C. Bantel. 2015. Assessing pain objectively: The use of physiological markers. *Anaesthesia* 70: 828–47.

Faggella, D. 2018a. The transhuman transition – Lotus eaters vs world eaters. May 27. https://danfaggella.com/the-transhuman-transition-lotus-eaters-vs-world-eaters/ (accessed October 7, 2019).

Faggella, D. 2018b. Programmatically generated everything (PGE). August 27. https://danfaggella.com/programmatically-generated-everything-pge/ (accessed October 7, 2019).

Koene, R. A. 2012. Embracing competitive balance: The case for substrate-independent minds and whole brain emulation. In *Singularity hypotheses*, ed. A. H. Eden, E. Steinhart, D. Pearce, and J. H. Moor, 241–67. Berlin and Heidelberg: Springer.

Loosemore, R. 2014. Qualia surfing. In *Intelligence unbound: The future of uploaded and machine minds*, ed. R. Blackford and D. Broderick, 231–39. Chichester: Wiley-Blackwell.

More, M. 2013. A letter to Mother Nature. In *The transhumanist reader: Classical and contemporary essays on the science, technology, and philosophy of the human future*, ed. M. More and N. Vita-More, 449–50. Chichester: Wiley-Blackwell.

Pearce, D. 2007. The abolitionist project.
https://www.abolitionist.com/ (accessed October 7, 2019).

Pitrat, J. 2013. *Artificial beings: The conscience of a conscious machine*. London: ISTE Ltd; Hoboken, NJ: John Wiley & Sons.

Sandberg, A., and N. Bostrom. 2008. *Whole brain emulation: A roadmap*. Technical Report #2008–3. Future of Humanity Institute, Oxford University.
https://www.fhi.ox.ac.uk/brain-emulation-roadmap-report.pdf (accessed October 7, 2019).

Skljarevski, V., and N. M. Ramadan. 2002. The nociceptive flexion reflex in humans – review article. *Pain* 96: 3–8.

Sloman, A. 1984. The structure of the space of possible minds. In *The mind and the machine: Philosophical aspects of Artificial Intelligence*, ed. S. B. Torrance, 73–82, Chichester: Ellis Horwood; New York: Halsted Press.

Sotala, K., and L. Gloor. 2017. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica* 41: 389–400.
https://www.informatica.si/index.php/informatica/article/viewFile/1877/1098
(accessed October 15, 2019).

Tegmark, M. 2017. Substrate independence. In *This idea is brilliant: Lost, overlooked, and underappreciated scientific concepts everyone should know*, ed. J. Brockman, 162–66. New York: HarperCollins.

Tomasik, B. 2019a. What are suffering subroutines? Updated May 17, 2019.
https://reducing-suffering.org/what-are-suffering-subroutines/ (accessed October 7, 2019).

Tomasik, B. 2019b. Do video-game characters matter morally? Updated June 14, 2019.
https://reducing-suffering.org/do-video-game-characters-matter-morally/
(accessed October 7, 2019).

Warpefelt, H., and H. Verhagen. 2015. Towards an updated typology of non-player character roles. In *Proceedings of the International Conference on Game and Entertainment Technologies*, ed. K. Blashki and Y. Xiao, 131–39. International Association for the Development of the Information Society.

Yampolskiy, R. V. 2015. The space of possible mind designs. In *International Conference on Artificial General Intelligence*, 218–27. Cham: Springer.

Yampolskiy, R. V. 2018. *Artificial Intelligence safety and security*. Boca Raton, FL: CRC Press.

Yampolskiy, R. V. 2019. Personal universes: A solution to the multi-agent value alignment problem. arXiv preprint arXiv:1901.01851.
https://arxiv.org/pdf/1901.01851.pdf (accesssed October 7, 2019).

Yampolskiy, R. V., and S. Ziesche. 2018. Preservation of personal identity – A survey of technological and philosophical scenarios. In *Death and anti-death, volume 16: Two hundred years after Frankenstein*, ed. C. Tandy, 345–74. Ann Arbor, MI: Ria University Press.

Yudkowsky, E. 2008. Artificial intelligence as a positive and negative factor in global risk. In *Global catastrophic risks*, ed. N. Bostrom and M. M. Ćirković, 308–45. New York and Oxford: Oxford University Press.

Ziesche, S., and R. V. Yampolskiy. 2019. Towards AI welfare science and policies. *Big Data and Cognitive Computing* 3(1)(Article #1). https://www.mdpi.com/2504-2289/3/1/2/htm (accessed October 7, 2019).