New Techno-Humanities 3 (2023) 41–48

Contents lists available at ScienceDirect

New Techno-Humanities

journal homepage: www.elsevier.com/locate/techum

# Artificially sentient beings: Moral, political, and legal issues

Fırat Akova

ARTICLE INFO

*Keywords:*
Artificial sentience
Artificial intelligence
Well-being
Hedonism
Aggregation

ABSTRACT

The emergence of artificially sentient beings raises moral, political, and legal issues that deserve scrutiny. First, it may be difficult to understand the well-being elements of artificially sentient beings and theories of well-being may have to be reconsidered. For instance, as a theory of well-being, hedonism may need to expand the meaning of happiness and suffering or it may run the risk of being irrelevant. Second, we may have to compare the claims of artificially sentient beings with the claims of humans. This calls for interspecies aggregation, which is a neglected form of interpersonal aggregation. Lastly, there are practical problems to address, such as whether to include artificially sentient beings in the political decision-making processes, whether to grant them a right to self-determination in digital worlds, and how to protect them from discrimination. Given these, the emergence of artificially sentient beings compels us to reevaluate the positions we typically hold.

## 1. Introduction

Artificially sentient beings can be defined as beings who have artificial sentience. While natural sentience is a sort of sentience that humans and non-human animals have through their biological substrates, artificial sentience is a sort of sentience that may be developed via technological means. To the best of our knowledge, artificially sentient beings do not exist yet, or, at least, their existence is undisclosed. According to some scholars, the question is not *whether*, but *when:* one of the most comprehensive literature reviews on artificial sentience finds that many scholars regard artificial sentience as possible.[1] Academic interest in artificial sentience is growing exponentially as the publications on artificial sentience between 2010 and 2020 vastly outnumber the publications on artificial sentience between 1990 and 2010.[2]

Rather than arguing for the non-zero probability of artificial sentience, I briefly summarize the methods which could pave the way for artificial sentience, and then explore the conceptual and practical questions its possibility raises.[3]

First, it may be difficult to understand the well-being elements of artificially sentient beings, measure their well-being, and compare it with the well-being of other sentient beings. I examine what implications this has for hedonism and show that hedonism may need to expand the very meaning of happiness and suffering, or, in some cases, it may run the risk of being irrelevant.

Second, we have to pin down the correct way of comparing the claims of artificially sentient beings with the claims of humans. For the most part, this calls for aggregating the claims of humans and aggregating the claims of artificially sentient beings separately. Here, interpersonal aggregation would amount to comparing the claims of different species, which means that we are in the territory of interspecies aggregation.[4] The case of artificially sentient beings could renew interest in interspecies aggregation, where there are many puzzles to address.

Lastly, there are issues pertinent to the political and legal status of artificially sentient beings. One issue is whether artificially sentient beings should have a right to partake in the political decision-making processes, which necessitates revisiting the boundary problem in democratic theory. Another issue is whether a right to self-determination could be granted to artificially sentient beings who live in digital worlds since the codes shaping their living conditions have to be controlled by someone or some group. The last issue considered is about protecting artificially sentient beings from discrimination, as they may be discriminated against based on their appearances and substrates.

[1] Jamie Harris and Jacy Reese Anthis, "The Moral Consideration of Artificial Entities: A Literature Review," *Science and Engineering Ethics* 27, no. 53 (2021): 2.

[2] Harris and Anthis, 6.

[3] Throughout the paper, I assume that artificially sentient beings deserve moral consideration by virtue of possessing sentience. There are attempts to justify giving moral consideration to artificial beings on the basis of sentience, consciousness, psycho-social properties, and living and information, to name a few. Consider Martin Gibert and Dominic Martin, "In search of the moral status of AI: why sentience is a strong argument," *AI & Society* 37, no. 1 (2021): 319-330; Kestutis Mosakas, "On the moral status of social robots: considering the consciousness criterion," *AI & Society* 36, no. 2 (2021): 429-443; Eric Schwitzgebel and Mara Garza, "A Defense of the Rights of Artificial Intelligences," *Midwest Studies in Philosophy* 39, no. 1 (2015): 98-119.

[4] As a subset of interpersonal aggregation, interspecies aggregation usually refers to a process where we aggregate the claims of humans and then aggregate the claims of non-human animals in order to compare them, though it more generally means aggregating and then comparing the claims of different species.

https://doi.org/10.1016/j.techum.2023.04.001
Received 4 April 2022; Received in revised form 27 January 2023; Accepted 26 April 2023

## 2. Artificial sentience as a non-zero probability

Theoretically, there are different paths to artificial sentience.

1. *Artificial enhancements* of *existing bodies and constructing artificial bodies from scratch*. Existing bodies, both human bodies and non-human animal bodies, can be enhanced.[5] Suppose that some people swap some components of their central nervous system, which are responsible for receiving sensations, with artificial components where they fulfil the same task. Instead of having natural components, they would have artificial components. A real-world example of an artificial enhancement is the artificial heart. Its transplant is currently possible, where it delivers the same tasks as the natural heart, and the lifespan of the patient is slightly extended. It is conceivable that just like an artificial heart, artificial components of the central nervous system, such as artificial neurons, can be developed to enhance existing bodies.[6] It is also imaginable that artificial bodies can be constructed without any natural components. If bodies are constructed from scratch without any natural components, or if existing bodies are greatly altered to the point that we are no longer certain to identify a being as a human or non-human animal, we may start to call that being an artificially sentient being.
2. *Whole brain emulation*. Whole brain emulation is often known as mind uploading.[7] The brain is scanned and modelled so that it can be copied to a digital world as a code. When the code is executed, it would result in running a simulation where it would act completely or almost like the original brain, which, in return, can produce experiences of sentience, consciousness, and intelligence.[8] Some artificially sentient beings of this sort can be called digital minds, who are "machine minds with conscious experiences, desires, and capacity for reasoning and autonomous decision-making".[9] In digital worlds, artificially sentient beings can either exist as a code without a body or they can be anthropomorphically designed like video game characters. The simulations can also generate their own digital worlds, which are called subroutines. There may be countless subroutines where artificially sentient beings could reside, as long as we have enough material, space, and energy.[10]
3. *Creating or finding new sources of sentience*. Humans are classified as carbon-based life forms. Carbons are the legacy, and the gifts, of events in space that occurred in the distant past, such as the Big Bang and the dissemination of stardust across the universe. What we refer to as biological is some combination of different materials including carbon, and sentience is a result of that combination. In the future, we may create new combinations, including or excluding carbon, which give rise to sentience. It is also possible that we come across new life forms possessing sentience in the universe, whose material combination does not consist of any material with which we are familiar.

As these paths to artificial sentience are speculative, some may be tempted to assign a tiny probability to the emergence of artificially sentient beings. Nonetheless, even a tiny probability should not prevent us from thinking about artificially sentient beings ahead of time. But why is pondering over artificially sentient beings important despite the fact that they are not here with us yet?

First, the expected value of artificially sentient lives can be enormous—even a tiny probability may result in extreme expected values, either positive or negative. To understand what this means, consider an example, existential risk. An existential risk is a type of risk that threatens to eliminate all sentient life on the Earth or at least threatens to radically curtail the long-term flourishing of our civilisation.[11] Since existential risks could bring about extreme suffering to many sentient beings, and result in the annihilation of the potential value of the future, some scholars argue that we must allocate significant resources to tackle them.[12] Even though they assign low probabilities to existential risks, they regard existential risk as very important because had it been happening, the expected value would be enormously negative. Something similar may happen with artificially sentient beings. Suppose that there is a one-in-a-million chance of the emergence of artificially sentient beings who are capable of reproducing themselves immensely fast, and, at the same time, for some reason, who would have net negative lives. They may be able to overwhelmingly outnumber humans, and even non-human animals. In that case, striving for eliminating the risk of such a dreadful scenario would be extremely important because zillions of artificially sentient beings could gravely suffer. Another low-probability

[5] Enhanced bodies are already imagined in the form of cyborgs. They reorient how we perceive bodies, encourage us to reconsider how we understand communication, and disrupt the assumed dichotomies between different bodies. For instance, David Gunkel notes that "In particular, the cyborg comprises a highly situated hybrid that does not adhere to the categorical distinctions by which the human subject would be distinguished and quarantined from its opposites. It is, therefore, a devious monstrosity that not only challenges the boundaries that had differentiated the human from the animal and the animal from the machine but also intentionally deforms the structure of all binary oppositions that construct and sustain Western epistemologies." in "We Are Borg: Cyborgs and the Subject of Communication," *Communication Theory* 10, no. 3 (2000): 347-348.

[6] Neil Harbisson, who is the co-founder of the Cyborg Foundation, identifies as a cyborg. Harbisson's skull has a permanently attached antenna. The antenna converts colours into sounds, and receives telephone calls and data from satellites. The case of Harbisson represents a small move towards artificially enhancing existing bodies, and many more enhancements may be possible. Another interesting case is the introduction of "teledildonics" which are devices allowing people to engage in sexual activity remotely. Such devices could be attached to one's body temporarily (or perhaps permanently in the future), where they may be controlled to trigger new experiences. For an analysis of teledildonics from a postphenomenological perspective, refer to Nicola Liberati, "Making out with the world and valuing relationships with humans," *Paladyn, Journal of Behavioral Robotics* 11, no. 1 (2020): 140-146.

[7] On the possibility of whole brain emulation, see Andreas Sanberg and Nick Bostrom, "Whole Brain Emulation: A Roadmap," Technical Report 2008-3, Future of Humanity Institute, University of Oxford.

[8] An interesting point of discussion here is whether having a body is a necessary condition for having sentience. While I do not aim to take a side here, I grant that positing that one can have sentience without a body (for instance, in a simulation where there are only codes) is more contentious than saying that one can have sentience through artificial enhancements to one's existing body. This not only opens up a discussion related to the nature of sentience but also related to whether machines only act as if they have sentience, and do not genuinely have sentience.

[9] Carl Shulman and Nick Bostrom, "Sharing the World with Digital Minds," in *Rethinking Moral Status*, eds. Steve Clarke, Hazem Zohny, and Julian Savulescu (Oxford: Oxford University Press, 2021), 306-326.

[10] Here, there is a distinction between *moral entities* and *moral agents*. According to Deborah G. Johnson, current computer systems lack mental states and intentionality, which bar them from being moral agents. With the whole brain emulation, it can be envisioned that artificially sentient beings will have both. Refer to Deborah G. Johnson, "Computer systems: Moral entities but not moral agents," *Ethics and Information Technology* 8, no. 4 (2006): 195-204.

[11] Existential risks include nuclear war, pandemic, runaway climate crisis, malign global governance, and insufficient control of artificial intelligence, to name a few. For a thorough examination of existential risks, refer to Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (New York: Hachette, 2020).

[12] The view that we have to allocate significant resources to prevent very harmful yet very tiny probability events or that we have to allocate significant resources to bring about very good yet very tiny probability events is sometimes called "fanaticism" or "Pascalian fanaticism." It is sometimes deemed a problem as it may have counterintuitive results. For a discussion on fanaticism, consider Nick Beckstead, "On the Overwhelming Importance of Shaping the Far Future," (PhD thesis, Rutgers University, 2013); Hayden Wilkinson, "In defense of fanaticism," *Ethics* 132, no. 2 (2022): 445-477; Hilary Greaves and William MacAskill, "The case for strong longtermism," GPI Working Paper No. 5-2021, Global Priorities Institute, University of Oxford, 24-26; Christian Tarsney, "The Epistemic Challenge to Longtermism," GPI Working Paper No. 10-2019, Global Priorities Institute, University of Oxford, 30-31.

scenario could be a conflict between artificially sentient beings and humans, which could result in net negative lives. Again, we have to eliminate that risk, and we can only do that by examining such scenarios.

Second, thinking on artificially sentient beings could help us to prepare ourselves for the upcoming moral, political, and legal problems arising from their existence. For instance, if artificially sentient beings would not be similar to humans, how can we understand the elements of their well-being? What should be our benchmark for allocating resources when there are competing claims between artificially sentient beings and other sentient beings, such as humans and non-human animals? Should artificially sentient beings enjoy the same legal rights as humans or should they be granted a subhuman status, where there are different legal implications? Is there a plausible and feasible way of politically representing artificially sentient beings? What should we do to adjust the social norms in protecting the interests of artificially sentient beings? These questions are just a few questions to which we have to have an answer.

Third, thinking about artificial sentience could encourage us to revisit our relationship with sentient beings other than humans, such as non-human animals. As humans can be speciesist and discriminate against non-human animals, we can take measures against spilling it over to artificially sentient beings, and, thanks to this, we may be more sensitive to discrimination at large.

Lastly, pondering over artificial sentience can enlarge our vision about what can happen in the future, and teach us to excel in longtermist thinking. After all, living with artificially sentient beings requires a long-term oriented strategy with a plethora of different predictions regarding how the future can be unlocked by artificial sentience.

In the following, I explore some of the above-mentioned issues in-depth.

## 3. New horizons of well-being

With the emergence of artificially sentient beings, new horizons of well-being could be discovered. What we understand from the very term well-being may change, as the well-being elements of artificially sentient beings may be slightly or dramatically different from the well-being of humans. We may even have a hard time making sense of what artificially sentient beings are feeling, or how and when their well-being is affected. These possibilities indicate new horizons of well-being.

As a theory of well-being, hedonism considers happiness and suffering as the only components of well-being.[13] I first list some of the possibilities regarding the well-being of artificially sentient beings below and then show the implications that they have for hedonism. I also include the possibilities which make no considerable difference in how we understand well-being and hedonism.

1. Artificially sentient beings may feel exactly like humans, or approximately like humans, where the well-being requirements of humans and the well-being requirements of artificially sentient beings would largely overlap.[14]
2. The happiness that artificially sentient beings receive or the suffering that they have to withstand may be totally different in sort, and we may not properly understand what they are feeling, though we could infer from our observations that they are receiving "happiness" or "suffering". This possibility is in line with what Carl Shulman and Nick Bostrom note about digital minds: "Bliss or misery more completely outside of the human experience might also be possible".[15]
3. Artificially sentient beings may face negative experiences which cannot be interpreted as suffering. Likewise, artificially sentient beings may benefit from positive experiences which cannot be traced back to happiness. They may have no capacity to experience happiness and suffering or their experiences may go beyond the duality of happiness and suffering.
4. Artificially sentient beings may experience astronomical states, including astronomical happiness and astronomical suffering, which may alter very meaning of happiness and suffering.

The first possibility is not so much different from the status quo. The hedonic range of artificially sentient beings is either identical or similar to humans, and, just like humans, they are capable of feeling happiness and suffering. Hedonism does not confront any problems in this possibility, besides the already existing objections to it.

In the second possibility, hedonism is still relevant to explaining the well-being of artificially sentient beings, as artificially sentient beings are capable of receiving happiness and suffering. However, since artificially sentient beings have different sorts of happiness and suffering that we do not experience or cannot grasp entirely, hedonists may be unable to single out each and every element that brings happiness and suffering to artificially sentient beings. This possibility calls for hedonists to widen their list of things that they think bring happiness and suffering to sentient beings, and, in this case, to artificially sentient beings.

The third possibility shows that hedonism could lose its relevance if artificially sentient beings lack happiness and suffering, or what artificially sentient beings are positively or negatively feeling cannot be classified as, reduced to or traced back to happiness and suffering, respectively. If artificially sentient beings only have feelings for which we have no words or concepts to describe, hedonists have two choices: they can abandon hedonism *merely* in case of explaining the well-being of artificially sentient beings and seek a new theory of well-being applicable to artificially sentient beings, or they can try to detect the well-being elements of artificially sentient beings that are assumedly the happiness and suffering equivalent of happiness and suffering that humans experience. In that case, at least two types of hedonism would appear: one for humans, and one for artificially sentient beings.[16] Indubitably, claiming that feelings X and Y that artificially sentient beings experience are equivalent to happiness and suffering that humans experience would be bitterly controversial.

Regarding the fourth possibility, imagine that there are artificially sentient beings who live in simulations. We can change the codes of simulations, which alter the living conditions of artificially sentient beings. For instance, by changing the codes, we can make a digital metropolis a digital desert, make every artificially sentient being laugh or cry at the same time, and target artificially sentient beings individually to adjust their experiences according to our wishes. In other words, we have total control over their lives. Further imagine that humans, Whitney and Billie, control two different simulations.

Whitney owns the first simulation. There is only one artificially sentient being in this simulation. Whitney wants to design a digital heaven for the artificially sentient being and develops a software such that it gives immense happiness to the artificially sentient being. The magnitude of this happiness is so great, so extreme, so unconventional that the happiness that the artificially sentient being has is beyond the comprehension of humans. The artificially sentient being feels astronomical happiness at the individual level. Whitney also decides to maxi-

[13] Ben Bramble, "A New Defense of Hedonism About Well-Being," *Ergo* 3, no. 4 (2016): 85.

[14] Sylvain Lavelle specifies two principles for understanding artificial sentience, one is strong, and the other is weak. Artificially sentient beings who feel exactly like humans are strong artificial sentience, whereas artificially sentient beings who feel approximately like humans are weak artificial sentience. Refer to Sylvain Lavelle, "The Machine with a Human Face: From Artificial Intelligence to Artificial Sentience," in *Advanced Information Systems Engineering Workshops*, eds. Sophie Dupuy-Chessa and Henderik A. Proper (Cham: Springer, 2020), 67.

[15] Shulman and Bostrom, "Sharing the World with Digital Minds," 311.

[16] It may be possible that there may be differences *among* artificially sentient beings themselves. In that case, we may have to generate new theories of well-being for each class of artificially sentient beings, and distinguish them from other theories of well-being.

mize happiness with the resources at stake, and by using another software, copies the artificially sentient being a million times. There are now a million artificially sentient beings in a digital heaven. Whitney has just brought about astronomical happiness at the collective level.

Billie owns the second digital environment. There is also one artificially sentient being in this simulation. Unlike Whitney, Billie is cruel and has evil aims. Billie wants to design the simulation as a digital hell. Billie runs a software that uninterruptedly tortures the artificially sentient being. The torture is so unbearable, so brutal, so detestable that the suffering the artificially sentient being receives is incomparable to any suffering that humans experience. The artificially sentient being feels astronomical suffering at the individual level. Billie has developed another software that copies the artificially sentient being a million times. There are now a million artificially sentient beings in a digital hell. Billie has just brought about astronomical suffering at the collective level.[17]

In the case of astronomical happiness and astronomical suffering, we may have to revisit the very meaning of happiness and suffering. One wonders whether we can still regard astronomical happiness as happiness. If it is still happiness, what sort of happiness is it? If it is not happiness, what is it? Likewise, does astronomical suffering still fall under the definition of suffering? If it is a misnomer, then what type of sensation is it? These questions are not only related to the boundaries of happiness and suffering, but also their scalability. We rarely think about the scalability of happiness and suffering, because we have biological limits to receiving them. But artificially sentient beings need not have such limits.

If astronomical happiness is not happiness but some other state, and, likewise, if astronomical suffering is not suffering but some other state, then hedonism is not relevant in cases where artificially sentient beings experience astronomical states. Once this is the case, the fourth possibility is subsumed under the third possibility, as we need new conceptual tools to explain what they are feeling. However, if we accept that astronomical happiness indeed falls under the umbrella of happiness, and, symmetrically, if we accept that astronomical suffering falls under the umbrella of suffering, then hedonism can still pinpoint happiness and suffering when explaining the well-being of artificially sentient beings, and hence remain intact.

Moreover, the possibility of artificially sentient beings who are capable of experiencing astronomical states, especially those who are able to experience astronomical happiness, raises the possibility of "utility monsters", which is a thought experiment by Robert Nozick.[18] Utility monsters have the ability to receive utility from a given unit of resource far greater than any other being. According to hedonistic utilitarianism, this may mean that the interests of others should be sacrificed to feed the utility monster—the utility monster is given priority in any resource distribution, where hedonistic utilitarianism becomes strictly inegalitarian. Artificially sentient beings who are capable of experiencing astronomical happiness can thus be utility monsters in the literal sense.[19] In that case, we may also have to reconsider the distinction between higher pleasures and lower pleasures, the plurality of valences, and resource allocation.

In all of these four possibilities, a discussion regarding what it means to feel happiness and suffering from the first person point of view emerges: the hard problem of consciousness. This traditional problem is now extended to artificially sentient beings. In discussing artificial suffering, Thomas Metzinger specifies four necessary conditions for the phenomenology of conscious suffering: (1) conscious experience (capability of having phenomenological states), (2) possession of a phenomenal self-model (the subjective experience that the person themselves are feeling something), (3) negative valence (frustrated preferences), and (4) transparency (the subjective certainty that one is feeling something, that one cannot be distanced from it).[20]

Metzinger argues that these conditions could apply to any kind of system, including artificially sentient beings. But there is an "epistemic indeterminacy" regarding artificially sentient beings. Epistemic determinacy means that "it is not the case that either we know that artificial consciousness *will* inevitably emerge at some point or we know that artificial consciousness will *never* be instantiated on machines. It is this neither-nor-ness that has to be dealt with in a rational, intellectually honest, and ethically sensitive way".[21] It seems that new horizons of well-being through the emergence of artificially sentient beings raise new questions regarding the very nature of well-being, how to understand consciousness, and the management of risk (for instance, the risk of vast amounts of suffering).[22]

## 4. Interspecies aggregation

When someone's interest is in conflict with someone else's, or when there is a trade-off between the gain of some group and the loss of some other group, we start to think about how to justifiably distribute benefits and burdens. There is a generic rescue case that is often used to illustrate the problems arising from comparing the claims of different individuals: suppose that we must choose between saving the life of one person against saving the lives of five people where we cannot save all of them. Some think that we should directly save the five people, because they constitute the greater number.[23] Some advocate tossing a coin to determine which group of people we should save, in which we would have assigned each individual an equal chance of being saved.[24] Some say that assigning probabilities of being saved according to numbers is the right way, that is, in this case, we should assign a 5/6 chance of being saved to the group of five people and a 1/6 chance of being saved to the group of one person.[25] Some others adopt a more nuanced position by comparing the value of utility (the number of people saved) with the disvalue of unfairness (the number of people being treated unfairly

[17] Note that I distinguish between astronomical suffering at the individual level and astronomical suffering at the collective level. The ordinary use of the term astronomical suffering refers to the collective experience and not to the individual experience. Digital worlds aside, astronomical suffering can arise from large-scale catastrophic wars, extreme harms from space colonisation, an unprecedented increase in animal farming and wild animal suffering, and unforeseeable developments that lead to horrific levels of suffering. Various institutions discuss astronomical suffering or aim to prevent it, such as the Future of Humanity Institute at the University of Oxford, the Center on Long-Term Risk, the Centre for Long-Term Resilience, and the Future of Life Institute. For some, superintelligence, a type of artificial intelligence whose intelligence far surpasses that of humans, can be a cause or cure for risks of astronomical suffering. Consider Kaj Sotala and Lukas Gloor, "Superintelligence as a Cause or Cure for Risks of Astronomical Suffering," *Informatica* 41 (2018): 389-400.

[18] Robert Nozick, *Anarchy, State, and Utopia* (Oxford: Blackwell, 1999), 42-43.

[19] A similar scenario has been noted by Shulman and Bostrom, and they are called "super-beneficiaries" in "Sharing the World with Digital Minds," 307.

[20] Thomas Metzinger, "Artificial Suffering," *Journal of Artificial Intelligence and Consciousness* 8, no. 1 (2021): 48-55.

[21] Metzinger, "Artificial Suffering," 47.

[22] For a thorough analysis of similar questions regarding artificial consciousness and artificial sentience, refer to David J. Chalmers, *Reality+: Virtual Worlds and the Problems of Philosophy* (New York: W. W. Norton & Company, 2022); Kenneth Einar Himma, "Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?" *Ethics and Information Technology* 11 (2009): 19-29; Thomas M. Powers, "On the Moral Agency of Computers," *Topoi* 32 (2013): 227-236.

[23] This is the standard utilitarian position.

[24] John Taurek thinks that tossing a coin "would seem to best express [our] equal concern and respect for each person." Refer to John M. Taurek, "Should the Numbers Count?" *Philosophy & Public Affairs* 6, no. 4 (1977): 303.

[25] Ben Saunders, "A Defence of Weighted Lotteries," *Ethical Theory and Moral Practice* 12, no. 3 (2009): 279-290.

by directly saving the greater number), which asks us to be sensitive to group sizes.[26] The territory in which these questions are debated is aggregation, specifically *interpersonal aggregation*.[27]

Scholars who work on aggregation have long focused on the cases of human beings versus human beings. The term interpersonal aggregation ordinarily refers to interpersonal aggregation within the same species—in other words, *intraspecies aggregation*. Nonetheless, there is almost no work on *interspecies aggregation*.[28] Interspecies aggregation would be concerned with trade-offs between different species. For instance, what should we do when we cannot save both five octopuses and one person? What should we do when we cannot save a hundred mantises, twenty octopuses, and one human at the same time? What should we do when we have to choose between saving one hawk and saving one lizard? To this date, the literature on aggregation has almost been human-only, and questions like the above, which recognize the diversity of species, have been neglected.

The emergence of artificially sentient beings would add a new layer of complexity to the already overlooked field of interspecies aggregation. Despite the fact that we are left with scarce literature on comparing the claims of humans and non-human animals, we would be asked to consider aggregation between artificially sentient beings and humans (and, in the long run, artificially sentient beings and non-human animals). How can we do that?

The first approach would be to treat the claims of humans equally with the claims of artificially sentient beings if their numbers are equal. The cases scrutinised by scholars in the literature of aggregation usually assume that there are no morally relevant differences between humans. It is easy to conceive such cases when humans are compared with other humans, as some differences between humans (for instance, skin colour, blood type, traits, etc.) can be assumed to be trivial when their fundamental claims are compared, such as their claims to continue living. But it may not be easy to conceive such cases when the claims of humans are compared with the claims of artificially sentient beings. There might be a wide range of different artificially sentient beings, such as those who physically live among us as robots, those who exist in digital worlds, and those who are based on a material that allows their sentience to be distributed across the whole universe. Assuming artificially sentient beings and humans would have no morally relevant differences would be very hard, as opposed to assuming some humans and some other humans have no morally relevant differences. For the first approach to be valid, artificially sentient beings have to be very similar to humans in terms of sentience, consciousness, intelligence, and so forth.

The second approach would be to aggregate the claims of different species separately and then weigh the claims of different species against each other. This is already done when the claims of humans are compared with the claims of non-human animals. Prone to objections, there is a widespread judgement that animals are less worthy than humans, and likewise, some animals are less worthy than other animals.[29] Similarly, some artificially sentient beings may be considered more worthy, or less worthy, than humans. What may be relevant in thinking either way? Their scale of sentience, psychological capacities, mental qualities, and some other features that are alien to us until we meet artificially sentient beings may all play a role.

A convenience that we have in aggregating the claims of some humans and comparing them with the claims of other humans is that humans evolve very slowly. For hundreds of thousands of years, we have had relatively specific and stable needs. What we call basic needs have not evolved much since the time of our earliest ancestors. Depending on the speed of the evolution of artificially sentient beings, we may need to change our weighing rapidly because some types of artificially sentient beings may be unprecedentedly fast in evolving.

The third approach would be to reject interspecies aggregation. There are at least two positions here.

The first position would be that we cannot aggregate the claims of sufficiently different species. For instance, a human's claim to having a minimally decent life and an artificially sentient being's claim to having a minimally decent life may wildly differ: for the former, a minimally decent life may require enough food, water, and security at the bare minimum, whereas, for the latter, a minimally decent life may require some amount of knowledge and access to some codes. Even if the term minimally decent life would be the same, the ways to satisfy the claims could be so different from each other that they ultimately become non-aggregatable when there are sufficiently different species. Imagine that we have two humans in a group, and against that group, we have another group that has one human and one digital mind. We can aggregate the claims of the first group, but it may be impossible to aggregate the claims of the second group, provided that humans and digital minds are sufficiently different species. In that case, one could reject interspecies aggregation when there are sufficiently different species.

The second position would be the standard anti-aggregationist stance, which rejects the plausibility of aggregation as a whole. Michael Otsuka spells out the Principle of Nonaggregation as the following: "one's duties to come to the aid of others are determined by the claims of individuals considered one by one rather than by any aggregation of the claims of individuals".[30] This means that we have to compare the claims of humans with the claims of artificially sentient beings one by one, rather than as a group. For instance, we would have to compare the claims of Anna, a human living in one of the traditional villages of Alaska, with the claims of Arc, an artificially sentient being who gets regular code updates to its simulation. According to anti-aggregationism, we cannot aggregate Anna's claims and the claims of some other humans, and compare them with the aggregated claims of Arc and some other artificially sentient beings. Anti-aggregationists already reject aggregation even when it is intraspecies, so it would be a surprise if they accept aggregation when it is interspecies.

## 5. Political and legal issues

A recent survey asks the participants the following question: "On a scale of 0–100, how much should your country's legal system protect the welfare (broadly understood as the rights, interests, and/or well-being) of the following groups?"[31] In the survey, 0 amounts to "not at

[26] Iwao Hirose, *Moral Aggregation* (New York: Oxford University, 2015); Martin Peterson, "Some Versions of the Number Problem Have No Solution," *Ethical Theory and Moral Practice* 13, no. 4 (2010): 439-451.

[27] Aggregation consists of interpersonal aggregation and intrapersonal aggregation. While interpersonal aggregation is concerned with evaluating the claims of different persons, intrapersonal aggregation is concerned with understanding the value of different temporal units of one's life. For ease of presentation, I use aggregation interchangeably with interpersonal aggregation.

[28] A notable exception is Shelly Kagan, *How to Count Animals, more or less* (New York: Oxford University Press, 2019).

[29] "On the one hand, many animals clearly do have some of the features that ground moral standing, so these animals *count*, morally speaking. Indeed, it is plausible to think that they count for far more than we ordinarily recognize. (Certainly they count for far, far more than one would think, given the appalling ways we normally treat them.) But at the same time, I think it is also clear that animals have *fewer* of the relevant features than people have (or they have them to a lesser degree), so that animals count for *less* than people. All of which is just to say: there are different degrees of moral status, and people have a higher status than that had by animals. What's more, and this is a third plausible implication of this basic line of thought, since animals themselves vary, one to the next, in terms of their possession of the relevant features, some animals have a higher moral status than others." in Kagan, 279.

[30] Michael Otsuka, "Skepticism about Saving the Greater Number," *Philosophy & Public Affairs* 32, no. 4 (2004): 415.

[31] Refer to Eric Martínez and Christoph Winter, "Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection," *Frontiers in Robotics and AI* 8 (2021): 2.

all" and 100 amounts to "as much as possible".[32] Artificially sentient beings are one of the groups targeted by the question.[33] The mean rating was 49.95% for artificially sentient beings.[34] This indicates that even at this stage where artificially sentient beings do not exist (or that their existence is undisclosed), there is significant support to protect them. Yet the perceived current level of the legal protection of artificially sentient beings is low, as the mean rating is 23.78%.[35] This means that there is a significant gap between the desired level of legal protection and the perceived level of legal protection. The survey asks the participants another question: "Insofar as the law should protect the rights, interests, and/or well-being of 'persons,' which of the following categories includes at least some 'persons?'."[36] For this question, 33.39% of participants lean towards or accept that at least some of the artificially sentient beings are persons.[37]

There is another survey, a census-balanced one, titled "Artificial Intelligence, Morality, and Sentience".[38] In this survey, 57.68% of participants agree that there should be a global ban on the development of sentience in robots/AIs. %37.16 of participants agree that we should grant legal rights to sentient robots/AIs. %30.31 of participants agree that the welfare of robots/AIs is one of the most important social issues in the world today.

There are other crucial questions that we may ask ourselves regarding the political and legal status of artificially sentient beings. One crucial question is whether artificially sentient beings have a right to partake in the political decision-making processes. There are several things to note here. First, so far, history has only witnessed the expansion of political rights within the same species—women, minorities and many ostracised groups earned their right to vote. But, for the first time in history, there is a non-negligible chance that politics may become interspecies with the inclusion of artificially sentient beings in politics. Second, responses to the boundary problem, which focuses on determining the scope of demos, may have to factor in new variables in deciding how to confer political rights. Third, some AI systems -although presumably not sentient- already take part in the political decision-making processes, such as in medicine, warfare, and automated vehicles.[39]

To answer the question at stake, we can appeal to the *all affected principle* and the *all subjected principle*, which are among the most discussed principles responding to the boundary problem. Note that with the arrival of artificially sentient beings, especially with the emergence of those who are living in simulations and subroutines, we may have to separate out the two spheres of politics: the first being the actual world that we are currently living in, and the second being the digital world that some of the artificially sentient beings could live in. The all affected principle and the all subjected principle could apply to both worlds.

The all affected principle states that those who are affected by a decision have a right to partake in the relevant decision-making process.[40] For instance, if one is affected by a change in tax laws, criminal codes or rules set by an authority, then one has a right to partake in the decision-making processes which alter them. Artificially sentient beings may be affected by many of the decisions that humans make. They may be affected by how resources are allocated (will they be able to get sufficient nutrition?), what investments are made (will they be able to increase their capacity?), which materials are produced (will they be able to repair themselves when damaged?), and which laws are enacted (will they be able to enjoy an adequate moral status?). Just as humans are affected by many of the decisions that other humans make in politics or everyday life, artificially sentient beings could be affected by them as well. If there are claims of artificially sentient beings in digital worlds which compete with the claims of sentient beings in the actual world, then they are all very likely to be affected by each other. For instance, the decision to determine how much energy should be allocated to digital worlds may be a field of competition, where all parties may be affected by the outcome. Artificially sentient beings may be affected by it because it may be a life-or-death situation for them, and the people living in the actual world may be affected because energy prices may change.

Nevertheless, there may be a gap between the actual world and the digital world. For instance, those living in the actual world may not be affected by whatever happens in simulations and subroutines, especially if the digital worlds at stake are self-sufficient. Digital worlds where some artificially sentient beings reside may involve suffering-free egalitarian utopias, extravagant lifestyles, large-scale conflicts over access to codes, or exploitative practices. But none of them may affect beings living in the actual world, may they be artificially sentient or not.[41] This resembles a situation where events happening in another galaxy do not affect the people living on Earth. In that case, according to the all affected principle, artificially sentient beings who live in digital worlds would have a right to partake in the decision-making processes in digital worlds only, but they would not have a right to partake in the decision-making processes in the actual world. Likewise, people who live in the actual world may only have a right to partake in the decision-making processes in the actual world, but they would not have a right to partake in the decision-making processes in the digital world.

The all affected principle is often contrasted with the all subjected principle, which is another principle determining who should have a right to partake in decision-making processes.[42] According to the all subjected principle, those who are subjected to a decision should have a right to partake in the decision-making process. If artificially sentient beings exist, they would likely be subjects of decisions. Think of artifi-

[32] Martínez and Winter, 2.

[33] Instead of "artificially sentient beings", the survey uses "sentient artificial intelligence", and the original question also targets many different groups other than artificially sentient beings.

[34] Martínez and Winter, 3.

[35] Martínez and Winter, 3.

[36] Martínez and Winter, 2.

[37] Martínez and Winter, 5.

[38] Janet Pauketat, Ali Ladak, Jamie Harris, and Jacy Anthis, "Artificial Intelligence, Morality, and Sentience (AIMS) 2021," *Mendeley Data*, V1, doi:10.17632/x5689yhv2n.1.

[39] Current artificial intelligence systems are not considered to be sentient. But it is claimed that they already partake in the decision-making processes. For instance, it is claimed that "We agree with the argument that AVs [automated vehicles] do not make decisions between the outright sacrificing of the lives of some, in order to preserve those of others. Instead, they decide implicitly about who is exposed to a greater risk of being sacrificed." in Maximilian Geisslinger, Franziska Poszler, Johannes Betz, Christoph Lütge, Markus Lienkamp, "Autonomous Driving Ethics: from Trolley Problem to Ethics of Risk," *Philosophy & Technology* 34, no. 4 (2021): 1042. Likewise, John P. Sullins claims that we are on the path to developing fully autonomous weapons systems to be used in warfare in "RoboWarfare: can robots be more ethical than humans on the battlefield?" *Ethics and Information Technology* 12, no. 3 (2010): 269.

[40] The all affected principle is one of the most used principles in determining the scope of demos. For a discussion of this principle, see Gustaf Arrhenius, "The Democratic Boundary Problem Reconsidered," *Ethics, Politics & Society* 1 (2018) 89-122; David Owen, "Constituting the polity, constituting the demos: on the place of the all affected interests principle in democratic theory and in resolving the democratic boundary problem," *Ethics & Global Politics* 5, no. 3 (2012): 129-152; Sofia Näsström, "The Challenge of the All-Affected Principle," *Political Studies* 59, no. 1 (2011): 116-134; Ben Saunders, "Defining the demos," *Politics, Philosophy & Economics* 11, no. 3 (2011): 280-301. For a more complex and refined view on the all affected principle, which distinguishes between the all actually affected interests principle, the all possibly affected interests principle, the all probably affected interests principle, and the all and only affected interests principle, consider Robert E. Goodin, "Enfranchising All Affected Interests, and Its Alternatives," *Philosophy & Public Affairs* 35, no. 1 (2017): 52-62.

[41] Remember that not all sentient beings would live in digital worlds. Robots who possess sentience may live among us in the actual world.

[42] For a discussion of the all subjected principle, consider Vuko Andrić, "Is the All-Subjected Principle Extensionally Adequate?" *Res Publica* 27, no. 3 (2021): 387-407; Andreas Bengtson, "Where Democracy Should Be: On the Site(s) of the All-Subjected Principle," *Res Publica* 28, no. 1 (2022): 69-84.

cially sentient beings in the form of robots physically living among us. They would not be able to escape from our laws—instead, we would expect them to abide by our laws governing all spheres of life. Once we expect them to abide by our laws and hold them liable if they do not, the all subjected principle asks us to grant them political rights, including a right to change those laws of which they are among the subjects.

What about digital worlds? Most probably, we would expect that the laws applicable to the actual world should also determine how artificially sentient beings living in digital worlds should live. Perhaps there could be additional laws governing digital worlds, as digital worlds may have specific features which cannot be regulated by laws created to regulate the actual world. Yet the all subjected principle would still apply here. Just like the all affected principle, the all subjected principle also applies to artificially sentient beings, endowing them with a right to partake in the decision-making process.

Departing from the all affected principle and the all subjected principle, there is a vital issue arising from the gap between the actual world and digital worlds: the question of self-determination.[43] Artificially sentient beings living in digital worlds would be quite vulnerable if the codes of digital worlds can only be changed by humans. To the extent that humans are free to create digital worlds as they please, artificially sentient beings living in digital worlds would be under serious threat. Their lives would single-handedly be shaped by the decisions of humans unless they have partial or full control over the codes. By handing the codes of digital worlds to their residents, a demand for justice may be satisfied as artificially sentient beings would be emancipated from unnecessary interference. To better understand how this makes sense, consider the following.

> *Residents*. A government decides to build a new city in a desert where there were no people in the past. It heavily invests in its design and infrastructure. Some people move there over time, populating the city. However, the residents of the city do not have any political rights over how the city should be run. For instance, there are no elections or referendums, and everything is decided through micromanagement by the government itself. Over centuries, the residents start to develop some form of distinct identity, and they become discontent over this authoritarian style of political management, as the government's wrong decisions make them vulnerable. Finding the current situation unjust, the residents demand some form of self-determination.

Now, think *Residents* along with *Digital Residents*.

> *Digital Residents*. A government decides to set up a new digital world where there were no artificially sentient beings in the past. It heavily invests in its design and code-writing process. Artificially sentient beings move from other digital worlds to this new digital world, populating it. However, artificially sentient beings who are now the residents of this new digital world do not have any political rights over how their digital world should be run. For instance, they have no access to codes through which some features of the digital world could be reworked, and all of the codes are retained, and changed if necessary, by the government itself. Over centuries, artificially sentient beings start to develop some form of distinct identity, and they become discontent over this authoritarian style of political management as the government's wrong decisions make them vulnerable. Finding the current situation unjust, artificially sentient beings demand some form of self-determination.[44]

There seems to be no morally relevant difference between the demand for self-determination in *Residents* and the demand for self-determination in *Digital Residents*. If we are sympathetic to granting self-determination in *Residents*, then we should also be sympathetic to granting self-determination in *Digital Residents*. Likewise, if we are not inclined to grant self-determination in *Residents*, then we should have no reason to do so in *Digital Residents*.

Exercising the right to self-determination in digital worlds does not only thwart unnecessary interference bringing about harm and vulnerability, but it also provides artificially sentient beings with the freedom to lead their lives according to their own preferences. This includes avoiding living in digital hells, and, perhaps more surprisingly, in digital heavens.[45] As opposed to perfected lives, some artificially sentient beings may attach importance to living "real" lives which include randomness and spontaneity. This may entail that one may desire not to live in digital heavens where what is going to happen is predetermined and life is linear. Moreover, regardless of whether digital heavens are predetermined, artificially sentient beings may want to experience some form and level of suffering, perhaps because of curiosity, or they may think that they have something to learn from that experience (such as to appreciate happiness more when confronted with suffering). Any legal and political framework recognizing the right to self-determination of artificially sentient beings should include the freedom to opt out from predetermined simulations and subroutines, as they may not be desirable *from the perspective of* artificially sentient beings.[46]

Building political and legal frameworks to protect artificially sentient beings is also important to avoid *substratism*. Sentience Institute defines substratism as "the unjustified disconsideration or treatment of beings whose algorithms are implemented on artificial (e.g. silicon-based) substrates rather than biological (i.e. carbon-based) substrates".[47] There are already some people who think that we would not be so different than artificially sentient beings, as we are just algorithms of some sort, and artificially sentient beings would be algorithms of some other sort: for instance, People for the Ethical Treatment of Reinforcement Learners consider humans as algorithms based on carbon, and, according to them, there may well be some other algorithms based on something else, such as silicon, to which we may owe moral consideration.

> We take the view that humans are just algorithms implemented on biological hardware. Machine intelligences have moral weight in the same way that humans and non-human animals do. There is

[43] Self-determination refers to one's ability to determine one's own destiny without unnecessary interference. It includes freely determining one's political status and freely cherishing the desired social, economic, and cultural values.

[44] The possibility that artificially sentient beings form a new identity raises an important discussion regarding whether artificially sentient beings can develop their identity as they wish or if their identity is pre-determined by the code that has been previously developed by the designer. While I do not aim to answer this question here, several authors have pondered over it. For instance, Dmytro Mykhailov suggests that intelligent decision-support system (IDSS) used in medicine can perform autonomous acts through deep learning mechanisms in "A moral analysis of intelligent decision-support systems in diagnostics through the lens of Luciano Floridi's information ethics," *Human Affairs* 31, no. 2 (2021): 149-164. Another example is "technological intentionality": Dmytro Mykhailov and Nicola Liberati argue that high-order programming languages such as C++ and unsupervised learning techniques like the "generative adversarial model" could display autonomous behaviour. A similar example is related to the "responsibility gap", where Andreas Matthias thinks that the manufacturer/operator cannot in principle predict the future behaviour of the learning machines based on neural networks, and machines could raise new behavioural patterns in "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and Information Technology* 6, no. 3 (2004): 175-183.

[45] The reasons against living in digital heavens, whatever they may be, share roots with the reasons not to plug in Robert Nozick's experience machine, which endows us with any pleasure we could ever desire.

[46] This line of thought breaks away from many utilitarian theories and sides with desire satisfaction theories.

[47] Jamie Harris, "The Importance of Artificial Sentience," Sentience Institute, February 26, 2021, sentienceinstitute.org/blog/the-importance-of-artificial-sentience. For Oscar Horta's work on speciesism, refer to Oscar Horta, "What is Speciesism?" *Journal of Agricultural and Environmental Ethics* 23, no. 2 (2010): 243-266.

no ethically justified reason to prioritize algorithms implemented on carbon over algorithms implemented on silicon.

The suffering of algorithms implemented on silicon is much harder for us to grasp than that of those implemented on carbon (such as humans), simply because we cannot witness their suffering. However, their suffering still matters, and the potential magnitude of this suffering is much greater given the increasing ubiquity of artificial intelligence.

Most reinforcement learners in operation today likely do not have significant moral weight, but this could very well change as AI research develops. In consideration of the moral weight of these future agents, we need ethical standards for the treatment of algorithms.[48]

Initially, artificially sentient beings are very likely to be discriminated against by humans on the basis of their substrates, just because they would not be recognised as having "real" sentience, where the word real is replaceable with "carbon-based".

Likewise, artificially sentient beings may suffer from *anthropomorphism*, which is the attribution of human characteristics to non-human entities. For instance, the interests of artificially sentient beings whose sentience system works more similarly to that of humans are likely to be considered more than other artificially sentient beings whose sentience system works less similarly to that of humans. As a result, some artificially sentient beings may be unjustifiably favoured while others are not unjustifiably disfavoured.

## 6. Conclusion

In this paper, I have explored various moral, political, and legal issues concerning artificially sentient beings.

First, I have analysed the new horizons of well-being, which may appear with the emergence of artificial sentience. By outlining four possibilities and scenarios regarding how sentience could arise in artificially sentient beings, I have demonstrated that understanding the well-being of artificially sentient beings may be quite hard to the point that we may have to revise or abandon a theory of well-being known as hedonism.

Secondly, I have maintained that we have to take interspecies aggregation seriously, and have shown some of the several directions that one may pursue in whether or how to aggregate the claims of artificially sentient beings. While the current literature often focuses on aggregating the claims of humans (and, to some extent, aggregating the claims of animals), I have demonstrated that the problems surrounding aggregation have to be revisited if artificially sentient beings start to exist at some point in the future. Aggregating the claims of artificially sentient beings would prove to be very complex, even more complex than aggregating the claims of humans.

Lastly, I have examined that political and legal issues surrounding artificial sentience, which deserve serious scrutiny. For instance, I have extended the all affected principle and the all subjected principle towards artificially sentient beings. To the best of my knowledge, these principles have never been scrutinised with regard to artificially sentient beings. The analysis has shown that artificially sentient beings can be captured by the all affected principle and the all subjected principle, which practically means that they may have a right to partake in the decision-making processes. I have also claimed that the right to self-determination could be extended to artificially sentient beings if they can create a distinct identity and political mechanisms independent of humans. I have also noted that we should be prepared for new forms of discrimination where artificially sentient beings may be discriminated against on the basis of their substrates or appearance.

As things stand, the emergence of artificially sentient beings is likely to propel us to review and revise the moral, political, and legal positions to which we ordinarily subscribe. If they ever come into existence, or if their existence is revealed, it is almost inevitable that artificially sentient beings will be the centre of our attention for a long time in the future.

## Declaration of Competing Interest

The author was supported by the Social and Environmental Entrepreneurs, Survival and Flourishing Projects. This paper does not necessarily represent the views of the funding body.

## Acknowledgements

I would like to express my gratitude to the Social and Environmental Entrepreneurs, Survival and Flourishing Projects for their support of this project. I thank the audience of the 10th Oxford Workshop on Global Priorities Research organised by the Global Priorities Institute, University of Oxford. I also thank the audience of the 96th Joint Session of the Aristotelian Society and the Mind Association hosted by the University of St Andrews. I lastly thank the audience of the 1st Analytic Philosophy Workshop organised by Poedat.

[48] "People for the Ethical Treatment of Reinforcement Learners," People for the Ethical Treatment of Reinforcement Learners, accessed November 19, 2021, petrl.org.