Taking Into Account Sentient Non-Humans in AI Ambitious Value Learning: Sentientist Coherent Extrapolated Volition

Adrià R. Moret

Independent Researcher/Currently unaffiliated
Barcelona, Spain
adriarodriguezmoret@gmail.com
ORCID: 0000-0002-2270-3730

(This is a draft, please cite the final version of the paper forthcoming or already published in the Journal of Artificial Intelligence and Consciousness)

Abstract: Ambitious value learning proposals to solve the AI alignment problem and avoid catastrophic outcomes from a possible future misaligned artificial superintelligence (such as Coherent Extrapolated Volition [CEV]) have focused on ensuring that an artificial superintelligence (ASI) would try to do what humans would want it to do. However, present and future sentient non-humans, such as non-human animals and possible future digital minds could also be affected by the ASI's behaviour in morally relevant ways. This paper puts forward Sentientist Coherent Extrapolated Volition, an alternative to CEV, that directly takes into account the interests of all sentient beings. This ambitious value learning proposal would significantly reduce the likelihood of risks of astronomical suffering from the ASI's behaviour, and thus we have very strong pro-tanto moral reasons in favour of implementing it instead of CEV. This fact is crucial in conducting an adequate cost-benefit analysis between different ambitious value learning proposals.

Keywords: The Alignment Problem · Coherent Extrapolated Volition · Suffering risk · Sentientism · Digital Minds · Non-human Animals

1. Introduction

The development of Artificial Superintelligence (ASI) (or Artificial General Intelligence poses serious existential and suffering risks (Bostrom, 2014; Yudkowsky, 2008; Sotala & Gloor, 2017; Tomasik, 2015a; Baumann, 2017a). To prevent such catastrophes, we must solve the alignment problem. We must ensure that (if we do so) the first ASI we create has certain values so that it does not result in those outcomes and so that it does that which we intend it to do. The value specification problem is part of the more general alignment problem (Gabriel, 2020; Christian, 2020). It is the question of what values to align an ASI with, and it is where this paper will focus instead of other more technical issues such as inner alignment.

Non-human sentient beings, such as domesticated non-human animals, animals living in the wild and possible future sentient digital minds, have usually been neglected in discussions about what values should be implemented into the first ASI to prevent catastrophic consequences. However, all sentient beings matter and the interests of non-human sentient beings should also be taken into account in any value learning proposal.

In this paper, I contend that in a future in which we could both implement some of the standard ambitious value learning proposals or alternative value learning proposals that directly took into consideration the interests of sentient non-humans, we would have both strong and *very* strong pro-tanto moral reasons to do the latter. However, in practice, as laid out in Section 5, it is not completely clear which kind of ambitious value learning proposal would be best to try to implement. I turn to the example of Coherent Extrapolated Volition (CEV) one of the most popular ambitious value-learning proposals and argue that it is not ideal since there is (at least) a sufficiently non-negligible chance that by not sufficiently including all sentient beings, it may lead to risks of astronomical suffering as a result of the own ASI's actions. Although here I only focus on CEV, similar arguments to the ones that I use here, are also applicable to many of the other ambitious value learning proposals out there and I may discuss and analyze other ambitious value learning proposals in future research.

2. Why the standardly proposed version of Coherent Extrapolated Volition is problematic

Coherent Extrapolated Volition (CEV) is Eliezer Yudkowsky's value learning proposal of implementing into an ASI the goal of fulfilling what we (humans) would agree we would want if given much longer to think about it, in more ideal circumstances (Bostrom, 2014). That is, what we (humans) would wish "if we knew more, thought faster, were more the people we wished we were, had grown up farther together; where the extrapolation converges rather than diverges, where our wishes cohere rather than interfere; extrapolated as we wish that extrapolated, interpreted as we wish that interpreted" (Yudkowsky, 2004). See (Yudkowsky, 2004) for a more detailed explanation of the ideal circumstances required for adequate coherent extrapolation. Some, including (Yudkowsky, 2016) may argue that this would, in fact, be enough for the ASI to adequately take into account the interests of non-human sentient beings if they ought to be taken into account. Similarly, it has also been contended that animals' well-being is also included in humans' preferences and that this together with the less myopic decision-making of the ASI (compared to human decision-making), might be enough (Russell, 2019). It has also been held that it might be difficult to sustain that humans should build an ASI that cares about the interests of sentient non-humans directly since then it would care more about them than what humans actually do (Russell, 2019). According to these views, if it is indeed the case that *if* humanity had more time to think about it in ideal circumstances, it would have the goal of adequately taking into account their interests, then by coherently extrapolating humanity's volition the ASI would take into account the interests of sentient non-humans. I will argue this answer is not sufficiently satisfactory for us not to have strong reasons to try to directly include the interests of sentient non-humans in the ASI's goals if we are able to do so.

It has plausibly been argued that for any value learning proposal, there are some aspects of the proposal that cannot be left to the ASI to be figured out, such as 'standing, concerning whose ethics views [or morally relevant interests] are included; measurement, concerning how their views [or morally relevant interests] are identified; and aggregation, concerning how individual views [or morally relevant interests] are combined to a single view that will guide AI behaviour' (Baum, 2020). As it has been argued, the alignment problem has an essentially normative component, different kinds of technical solutions could be used for loading certain values into the reward function of a future A(S)I likely have different implications on exactly what kind of values we can implement, and thus, "there is no way to 'bracket out' normative questions altogether" (Gabriel, 2020). Ultimately, the designers of the A(G)I that may become an ASI, cannot abstain from facing ethical decisions regarding in what way, what and whose values or morally relevant interests they include in the ASI's goals. I contend that we should take seriously the possibility that Yudkowsky's popular proposal does not adequately take into account moral considerations pertaining to the moral considerability of sentient non-humans.

Even though the standard CEV proposal adequately avoids directly specifying the values that should guide the ASI's behaviour (and as a consequence plausibly avoids substantial value lock-in), it nonetheless specifies which sentient beings have their volitions coherently extrapolated. The standard proposal only includes presently existing humans. In this sense, the standard CEV proposal excludes all other sentient beings and gives a specific answer to the standing problem. I contend that there are very strong reasons to suggest that CEV's standardly proposed solution to the standing problem is much more morally undesirable than an ambitious value learning proposal that takes the interests of all sentient beings more into account. From now forward I will refer to this kind of alternative proposal that I will further develop in Section 3 as *Sentientist Coherent Extrapolated Volition (SCEV)*.

CEV excludes the volitions of sentient non-humans from the extrapolation base, in section 2.1. I will show why their mere exclusion is unjustified in the same way that it would be unjustified to exclude groups of humans. This in itself gives us strong moral reasons to implement SCEV instead of CEV if we are able to do so. However, in section 2.2. I contend that we have *very* strong moral reasons to do so. There is (at least) a non-negligible probability that an adequate implementation of the standard CEV proposal results in the ASI causing or allowing the occurrence of risks of astronomical suffering (s-risks). These are risks of events that bring about suffering in cosmically significant amounts, vastly exceeding all suffering that has existed on Earth so far (Sotala & Gloor, 2017; Althaus & Gloor, 2016). "Significant" here means relative to expected future suffering (Althaus & Gloor, 2016). Taking the badness of s-risks into account, this non-negligible probability is sufficiently large for CEV to be a much more undesirable solution to the value specification problem than SCEV.

It is important to note and keep in mind that the reasons in favour of SCEV that I present and discuss in this section and throughout this paper are pro-tanto in the sense that (as discussed in Section 5) they could be overridden or defeated by other pro-tanto reasons against the implementation of SCEV. As a result, it may not be clear whether trying to implement SCEV instead of CEV is the right choice all things considered.

2.1. Why we would have strong moral reasons to implement SCEV instead of CEV: an analogy with CEO-CEV and men-CEV

To see why it is unjustified for sentient non-humans to be excluded from the extrapolation base, we can look at different alternative (and fictional) CEV proposals. Take, for example, the *CEO-CEV* proposal: the proposal of only applying CEV, that is, of only coherently extrapolating the volition, of the CEO of the research lab that first develops ASI or AGI. Or, for example, take the *men-CEV* proposal of only coherently extrapolating men's volitions. These would clearly be substantially more morally undesirable ambitious value learning proposals to employ in comparison to CEV if we could choose to implement any one of them. We have strong reasons against only coherently extrapolating the volitions of a small subset of humans that is unrepresentative of humanity as a whole.

Probably the most plausible reason against pursuing such paths as ambitious value learning proposals (if the possibility of implementing normal CEV is present) is that there is clearly (at least) a sufficiently large non-negligible probability that these proposals would result in the ASI's behaviour having negative consequences for the humans not directly included in the extrapolation base of the CEV. People who are not men or people who have sufficiently different preferences from the AI lab CEO's preferences could find themselves in a post-ASI world where their interests are not sufficiently taken into account, and as a result, suffer negative consequences. Their interests could be very neglected as a result of only being present in the ASI's utility function to the extent to which men or the AI lab's CEO cared about them once they thought enough about them in ideal circumstances.

Even though it is possible that if the AI lab's CEO or men had sufficient time to think under the ideal conditions, then, they would want to take equally into account the interests of the people outside the extrapolation base as much as theirs, we still would not accept these proposals as adequate proposals for ambitions value learning compared to normal CEV. Because there is a sufficient likelihood that this possibility does not turn out to be the case. Whilst these value learning proposals might plausibly be better than not implementing any values into the ASI, they are significantly worse than CEV because there is a sufficiently large non-negligible probability that they would result in substantial negative consequences for the people excluded from the extrapolation base in each case. This would give us strong reasons to rule out CEO-CEV and men-CEV in a decision-making situation where we could also choose to implement the normal CEV proposal that takes all humans into account.

I contend that for similar kinds of reasons, CEV is also an inadequate proposal in comparison to SCEV. The normal CEV proposal, like CEO-CEV and men-CEV, excludes a subset of moral patients from the extrapolation base. It excludes all sentient non-human beings. And as in the previous examples, if CEV were to be adequately implemented, the interests of sentient non-humans would only be in the ASI's utility function and would only be taken into account to the extent to which humans cared about them once they thought enough about them in ideal circumstances. Sentient non-humans are moral patients, they could be affected in morally undesirable ways, as could the people excluded from the extrapolation base in the previous examples. This could occur as a mere result of their interests not being sufficiently important to the individuals in the extrapolation base even if they had enough time to think about it under ideal circumstances. This is why if we were able to do so, instead of implementing CEV, we would have strong moral reasons to implement SCEV: a proposal that took the interests of all sentient beings adequately into account. Below, in Section 4, I address two possible objections that contend that there is a relevant difference between excluding other humans and excluding sentient non-humans.

2.2. *Why we would have very strong moral reasons to implement SCEV instead of CEV: significantly reducing risks of astronomical suffering*

We do not only have *strong* moral reasons to implement SCEV instead of CEV. There are further reasons to believe that the difference in expected negative consequences between implementing CEV instead of SCEV is much larger than the difference in expected negative consequences between implementing CEO-CEV or men-CEV instead of CEV. If we could do so, we would have much more pressing reasons to aim for SCEV instead of the standard CEV proposal, than the reasons we have to aim for the standard CEV proposal instead of CEO-CEV or men-CEV. Then, if we could do so, we would have *very* strong moral reasons to implement SCEV instead of CEV.

2.2.1. The greater lack of moral consideration for sentient non-humans

It is likely that any given group of humans (without any particular psychosocial conditions) cared much more about the interests of the rest of humans if they had enough time to think about it under ideal circumstances than, the extent to which all humans would care about the interests of all sentient non-humans if they had enough time to think about it under ideal circumstances. It is likely that in the first case the humans in the extrapolation base granted much more consideration to the humans outside the extrapolation base, than the amount of consideration that, in the second case, all humans would grant to sentient non-humans outside the extrapolation base.

Of course, we cannot be sure that is the case since we do not specifically know how coherently extrapolating the volitions of any subset of or all humans would work in practice. And we do not know with certainty what any subset of or all humans would agree that they want, if given much longer to think about it, in more ideal circumstances. However, it is not the case that we can say nothing about which outcomes from different types of CEV are more or less likely. Indeed, we can say something. For instance, humans tend to have much more consideration for those who are closest and more similar to them (Caviola et al., 2018; Dhont, Hodson, & Leite, 2016), and this may affect the process of coherently extrapolating volitions in a negative manner for sentient non-humans. This is due to the fact that sentient non-humans, such as domesticated non-human animals, animals living in the wild and possible future digital minds are much more distant (in a broad social, emotional, cognitive, psychological and even spacial and temporal sense) to humans than other humans.

Furthermore, we should also update on the fact that (at least partially) as a result of not equally taking into consideration the interests of all sentient beings, currently and in the past, humans have killed and caused immense amounts of unbearable suffering to an immense number of non-human sentient beings for the almost negligible benefit of taste pleasure. To see how much we should update on this fact, imagine possible worlds in which the AI lab's CEO or men caused immense amounts of suffering to the humans excluded from the extrapolation base in each case. And did so in proportional quantities to the amount of suffering that humans in this world cause to other animals. Imagine they did this via, for instance, torturing the humans outside the extrapolation base in each case. Imagine, that, as humans do to factory-famed non-human animals, they caused unbearable suffering to the humans outside the extrapolation base throughout their lives and killed almost all of them before they could even reach adulthood. In these possible worlds, we would have more reasons to believe that it is more likely that implementing CEO-CEV or men-CEV would lead to worse consequences than the ones we expect it to have in the actual world. Analogously, it seems plausible that it is significantly more likely that in this world, in a world where human psychology has (at least in part) led humans to commit atrocities, such as factory farming, to other non-human sentient beings, it is also the case that the standard proposal of CEV leads to more negative consequences for non-human sentient beings than in a world in which our human psychologies had not played a role in committing atrocities towards other sentient beings.

2.2.2. The nature and different kinds of the worst outcomes: risks of astronomical suffering

As we have seen there is (at least) a non-negligible chance that the interests of sentient non-humans would be highly disregarded by an ASI with the implementation of the standard CEV proposal. And that the likelihood of this (whatever it is), is higher than the likelihood that the interests of the humans not included in the extrapolation base are highly neglected in an ASI-controlled future where only the volitions of some humans are coherently extrapolated. A civilization with an ASI with either of those kinds of ambitious value learning proposals is (at least somewhat) likely to want to expand and settle space and make use of advanced AI and simulation technologies. In such a future where the interests of sentient non-humans would be highly disregarded, there is a significant probability that sentient non-humans would suffer astronomically as a result of the actions the humans would want to take and would likely be taken by the ASI in following their CEV.

Wild animal suffering might be spread through the universe by such civilisations. This could be a result of direct panspermia or intentionally terraforming planets because of some of the humans' values to spread life through the universe or for their aesthetic enjoyment of nature (Tomasik, 2018; O'Brien, 2021). But if no preventive measures are put in place it may also occur as a result of not adequately decontaminating spaceships, which could result in some microbes or other living organisms being transported accidentally to other planets which could over the course of many years result in the emergence of sentient lifeforms in those planets. Since it is plausible that in the lives of these future non-human animals suffering would be very prevalent and possibly even predominate, as in most of the lives of currently existing non-human animals living in nature (Ng, 1995; Groff & Ng, 2019; Tomasik, 2020; O'Brien, 2021) if this were to occur it could lead to astronomical amounts of suffering.

A civilization with an ASI that takes into account the interests of some or all humans but not those of sentient non-humans might also make abundant use of advanced AI and simulation technologies. This could either knowingly or accidentally result in astronomical amounts of suffering. An ASI that neglects the interests of artificial forms of sentience might accidentally create vast quantities of sentient subroutines. These are instrumentally valuable functions, subprograms, robot overseers, robot scientists or subagents inside the ASI's program and structure (Tomasik, 2019a). In the same way that the emergence of phenomenal experience and the capacity for suffering has plausibly been evolutionary instrumentally useful in guiding behaviour, it might be instrumentally useful for the ASI in controlling the behaviour of its internal process in order to achieve its goals. By undergoing a complex optimization process, the ASI might instrumentally value the creation of suffering as has done natural selection, a complex process optimizing for gene reproduction (Sotala & Gloor, 2017).

Furthermore, an ASI-controlled civilisation where CEV is implemented and the interests of sentient non-humans are highly disregarded might also result in Mind Crime, the creation of vast quantities of digital minds in simulations, including suffering ones. It might create immense numbers of ancestor simulations of past human history, of natural environments or of possible future evolutionary paths for research purposes (Bostrom, 2014, pp. 125-26; Sotala & Gloor, 2017). In many of these, if the simulations are sufficiently detailed and complex, "human" digital minds and digital minds of both wild or domesticated non-human animals might emerge thus multiplying wild animal suffering and the subjective experiences of being factory-farmed (Tomasik, 2015b; Baumann, 2017a). It could also intentionally create or allow other humans to intentionally create simulations with a specially significant amount of suffering, for entertainment or to satisfy sadistic preferences.

In wanting to explore all the space of possible minds it might also try to simulate various different distant, weird and alien forms of sentience compared to ours (Tomasik, 2015b). The possibility that any of these kinds of non-human sentient beings came to exist on immense scales and that their interests were neglected by the CEV-aligned ASI in power aggravates the risk of astronomical suffering.

And, as it has very plausibly been argued elsewhere, 'nearly all plausible value systems will want to avoid suffering risks and for many value systems, suffering risks are some of the worst possible outcomes and thus some of the most important to avoid' (Sotala and Gloor, 2017). On any value system that holds that it is very important to prevent undesired and intense suffering, it will be of extreme importance to prevent astronomical amounts of undesired and intense suffering. Then, it need not be the case that the likelihood of these outcomes is very high, for it to be of immense importance to prevent them. If there is a non-negligible or higher chance of the occurrence of s-risks it is still of great moral importance to reduce it since their occurrence would be extremely morally undesirable.

2.2.3. Why (even without taking into account the higher disregard for their interests) sentient non-humans would more likely be the victims of s-risks

Due to differences between the kinds of beings that biological humans and sentient non-humans are, even if it were not the case that sentient non-humans are more likely to be disregarded, it would still be significantly more likely that sentient non-humans would be harmed much more by implementing CEV that disregarded their interests, than excluded humans by implement CEO-CEV or men-CEV that disregarded their interests. This is so because it is easier to cause astronomical suffering to sentient non-humans than it is to cause astronomical suffering to excluded (biological) humans.

An ASI can cause actions that result in astronomical suffering or it can allow actions by other agents that cause astronomical suffering. In both cases, whether the action is performed by the ASI or by other agents, the fact that the action results in astronomical suffering can be intentional to a greater or lesser extent. When the actions that result in astronomical suffering are maximally intentional, the agent that performs them wants to cause astronomical suffering, when they are maximally unintentional, the agent does not even know that they are causing astronomical suffering. In the middle areas of this scale of the intentionality and unintentionally of the action, there are cases in which the agent knows that they are causing astronomical suffering but acts anyway in order to achieve its goals, and there are cases in which the agent is uncertain about whether they are causing astronomical suffering but also acts anyway. There are a few characteristics that many kinds of sentient non-humans have which make it more likely and easier for them to endure astronomical suffering. Biological humans excluded in CEO-CEV or men-CEV lack these characteristics, and thus it is more difficult for them to be the victims of astronomical suffering. Here are some of the relevant differences:

- *Resource efficiency of reproduction or replication:* Plausably much fewer resources are required for suffering digital minds to be copied relative to the resources required to sustain them than the resources required for suffering humans to reproduce relative to the resources required to sustain them (Sotala, 2012; Shulman & Bostrom 2021). This makes the agent causing the astronomical suffering more likely to intentionally replicate digital minds and more easy for them to do so unintentionally, than with (biological) humans.

- *Lack of intervention in sustaining a suffering population:* Plausablly less intervention is required to sustain a suffering population of digital minds or (biological) non-human animals living in nature than a suffering population of (biological) humans. In the former case, the agent causing the astronomical suffering can just let the suffering continue indefinitely (either more intentionally or more unintentionally). In the latter case, the processes of natural selection, predation, and r-selection in reproductive strategies continue without supervision and lack of intervention (either more intentionally or more unintentionally). However, more intervention is required in the biological human case to prevent them from trying to take control over that which causes them suffering, prevent them from trying to use resources to decrease the disvalue in their lives and increase the value in it and prevent them from trying to cease to exist via ceasing to reproduce or by suicide.

- *Uncertainty about the presence of suffering in some cases:* In the mentioned case in which the agent is uncertain about whether they are causing astronomical suffering but also acts anyway, they may not choose to act in such a way if they knew that they were causing astronomical suffering. It is (and might probably continue to be) much more unclear whether *certain* animals living in nature or *certain* possible digital minds are enduring suffering than whether a given (biological) human is enduring suffering. The latter one may be able to communicate this fact and resembles us much more, and thus

> would be less likely to be the victims in the cases the agent causing the suffering is uncertain about whether they are doing so.

All these factors together with the fact that (biological) humans would be less likely to be disregarded make it substantially less likely that they would be the victims of s-risks compared to the likelihood that sentient non-humans would be such victims in a future controlled by an ASI with an ambitious value leaning proposal that did not directly coherently extrapolate their volitions.

In conclusion, there are two ways of justifying why we would have pressing reasons to implement an ambitious value learning proposal such as SCEV instead of the standard CEV proposal if we were able to do so. On the one hand, in the same way that there are strong moral reasons to implement CEV instead of a proposal for ambitious value learning such as CEO-CEV or men-CEV, there are also strong moral reasons to implement SCEV instead of CEV. On the other hand, since there is (at least) a non-negligible chance that the interests of sentient non-humans are highly disregarded by an ASI with the standard CEV proposal, and in such a future where the interests of sentient non-humans are highly disregarded it is substantially likely that sentient non-humans suffer s-risks, there is (at least) a non-negligible chance that the implementation of the standard CEV proposal results in s-risks for sentient non-humans from the own ASI's actions. Since s-risks are plausibly the worst possible outcomes and extremely morally undesirable under plausible normative views, even a (at least) non-negligible chance of their occurrence is very bad. Then, the standard CEV proposal that excludes the immense majority of present and future sentient beings is very morally undesirable if a better alternative is available. This increased likelihood of s-risks from the ASI's behaviour due to implementing certain kinds of ambitious value learning (as I shall argue in the following section) can be significantly reduced by implementing a truly sentientist solution to the value specification problem.

3. An alternative sentientist proposal for ambitious value learning

Here, I shall present Sentientist Coherent Extrapolated Volition, an ambitious value learning alternative to CEV. This proposal consists of coherently extrapolating the volitions of all currently existing and future sentient beings. That is, to include all moral patients in the extrapolation base of the CEV.

> *Sentientist Coherent Extrapolated Volition*: the goal of fulfilling what all (affectable i.e. present and future) sentient beings would agree that they want, if given much longer to think about it, in more ideal circumstances.

Even though this is very ambitious, the standard CEV proposal also is, even if it is a bit less so. And if the standard CEV proposal is in-principle doable, I see no reason to believe that this proposal is not in-principle doable as well. For the standard CEV proposal to be realized, it would have to be determined what humans' desires would look like if we knew more, thought faster, were more the people we wished we were and had grown up farther together among other things (Yudkowsky, 2014). This would require determining what we would want if we had different capabilities than the ones we currently have. It would require determining what we would value if we had superhuman capabilities in many respects, such as, among other things, not being affected by cognitive biases, and being able to process and comprehend more amounts and different kinds of information (Yudkowsky, 2014).

Concretely, this would or could consist of (among other things) creating a "detailed model of each individual mind, in as much detail as necessary to guess the class of volition-extrapolating transformations defined by "knew more," "thought faster," etc." (Yudkowsky, 2004). The objective would be to approximate and understand the kinds of idealised changes that each mind would undergo if volition-extrapolating upgrades (such as knowing more, and having more time to think) were performed on it. It is important to understand that the proposal is not for these volition-extrapolating upgrades to be directly performed by the ASI on the minds of the individuals included in the extrapolation base by physically intervening in their brains. Instead, the proposal would be to perform these changes on detailed models of the minds of all the individuals in the extrapolation base which would be generated by the ASI. Neither CEV nor SCEV proposes that the volition-extrapolating upgrades are directly performed on the minds of the individuals in the extrapolation base. This would just result in all of them (humans and non-humans) being suddenly and directly uplifted and enhanced to similar levels of cognitive capability as that of the ASI. Once the ASI had generated sufficiently detailed models of each individual mind, and had applied volition-extrapolating upgrades to each of these models, their resulting coherent preferences would be the ones that would guide the ASI's behaviour.

If with more advanced technology, these specific processes are in principle doable and understandable in human minds it is likely that they could also be done with more simple minds such as those of many non-human animals. And, it is also plausible that we could at least have a more-than-zero probabilistic approximation of what these idealized changes would look like for possible future digital minds after having undergone volition-extrapolating upgrades. Then, it seems likely that if the process required for coherently extrapolating the volitions of humans is theoretically possible, it is also so with the volitions of sentient non-humans. And even if we could not in practice reach a fully idealized state of extrapolation or a state of full coherence with and between (both human and non-human) volitions, what is relevant is that we get as close as we can to achieving these. Doing so would be much more preferable than a proposal that did not result in any coherence between volitions, and that did not extrapolate the volitions at all.

All present and future sentient beings, including humans, both domesticated and non-domesticated non-human animals and possible future digital minds, have goals by the mere fact of being sentient. At least, in the broadest sense possible, they have desires and preferences to have positive phenomenal experiences and be free from negative ones. And many of them might have goals much richer and more complex than these. All of these goals form part of each of their volitions, which could and should be coherently extrapolated. They could in principle be extrapolated as human volitions can even if many sentient non-humans possess certain cognitive capacities to a lesser degree than humans usually do. This is so because as with human volitions, we could also (among other things) determine the changes that different kinds of sentient non-human minds undergo when volition-extrapolating transformations are performed on them. Even if many sentient non-humans do not possess the capacity to do certain types of abstract thinking about their goals or reflect on their values, humans also do not possess the capacities necessary to determine the content of their extrapolated volitions. In both cases, volition-extrapolating transformations would enhance the capacities of both sentient humans and non-humans to adequately determine the content of their extrapolated volitions.

In modelling what changes in preferences would occur if volition-extrapolating transformations were applied to the volitions of existing non-human animals, such as squirrels, SCEV may arrive at models of extrapolated volitions with weird preferences. We currently have no idea what they would look like. However, this is not enough to suggest that we should not include squirrels or other non-humans in the extrapolation base. In fact, it has been argued that there are good reasons to uplift the (cognitive) capacities of non-human animals (at least) to similar levels than the ones had by most humans for democratic reasons (Paez & Magaña, 2023), for Rawlsian liberal-egalitarian reasons (Dvorsky 2008), and anti-speciesist reasons (Chan, 2009). And, as I will argue in Section 4.3. We have no good reasons to believe that implementing SCEV or enhancing their capabilities would cause them to cease to exist, but rather, can be done while preserving their identity (Paez & Magaña, 2023). Furthermore, as I show in Section 4.1. it is implausible to believe that including non-human animals in the extrapolation base could result in us being harmed by the weird preferences their extrapolated volitions might have. And thus, since, it remains the case that by not including them we are likely increasing the risk of astronomical suffering, we have strong reasons to do so even if we are highly uncertain about the content of their extrapolated volitions.

3.1. A solution to the standing problem

Since, as mentioned above, we cannot leave the solution to the standing problem (the issue of whose interests or volitions are included in the extrapolation base) to be figured out by the ASI, we should take into account uncertainty when making such a decision ourselves. What is relevant in determining what entities should morally be taken into account, and thus, also be included in some way or another in the extrapolation base, is not whether or not we can be sure that they possess the relevant traits for moral patienthood. Against what Yudkowsky has contended, the fact that we could be wrong about whether a given entity deserves moral consideration is not a sufficient reason to exclude them from our moral consideration or form the extrapolation base (Yudkowsky, 2016). Instead, what is relevant is whether there is a non-negligible chance that they possess the relevant traits for moral patienthood. This is so because if there is some chance that a given entity can be harmed in morally relevant ways by our actions, it is wrong to perform those actions all else equal. Because some chance of harm is worse than no harm. Thus, what is relevant in determining which entities should be morally taken into account is whether there is a non-negligible chance that they are sentient (Sebo, 2018). Furthermore, as we have seen, being included or not in the extrapolation base can indeed directly affect the extent to which (in expectation) the interests of given beings are respected and taken into account by the ASI's behaviour. Then, all entities for which there is a non-negligible chance that they are sentient should be included in the extrapolation base. We have very strong reasons to believe that there is at least a non-negligible probability that almost all present and future humans, non-human animals and possible future digital minds are moral patients, and thus should directly be included in the extrapolation base. This is so because of three main reasons. First, there is a wide consensus and strong evidence on the sentience of vertebrate animals and many invertebrate animals (Low, 2012; Proctor et al., 2013; The Universal Declaration on Animal Welfare, 2007; Waldhorn, 2019a; Waldhorn, 2019b), and there is wide agreement that artificial entities, such as digital minds, with the capacity for sentience will or could be created in the future (Harris & Anthis, 2021). Secondly, sentience in the sense understood above as the capacity of having positively or negatively valenced phenomenally conscious experiences is widely regarded and accepted as a sufficient condition for moral patienthood (Clarke, S., Zohny, H. & Savulescu, J., 2021). And thirdly, it is also the case that it seems very plausible and it is also widely accepted that the intrinsic value of good things or bad things or the intrinsic moral importance of respecting and considering the interests of moral patients cannot be discounted solely because they occur in the future, see (Greaves & MacAskill, 2021: p.18; Ord, 2020: p.52; Beckstead, 2013: p.18) for acceptance of this view in the literature. Then, all present and future entities for which there is (at least) a non-negligible chance that they are sentient should directly be included in the extrapolation base.

3.2. How to include the extrapolated volitions of future not-yet-existing sentient beings

According to SCEV, apart from currently existing sentient beings, all future sentient beings should have their volitions directly included in the extrapolation base, since then the probability of s-risks from the ASI's actions would be significantly reduced. But, how exactly should the volitions of future not-yet-existing sentient beings be directly included in the extrapolation base? At any given point in time *t*, the ASI should take those actions that would in expectation most fulfil the coherent extrapolated volition of all sentient beings that exist in *t*. It is the case that many (or almost all) actions that the ASI could or would take would change the kinds and number of sentient beings that would exist after it had performed the given action(s). Due to this fact, after any given action the ASI would have to incorporate the extrapolated volitions of the new currently existing sentient beings if there are any, and then decide how to act again based on the coherent extrapolated volitions of all existing beings at that point in time.

It is important to realize that other kinds of proposals about how to take into account future not-yet-existing sentient beings, might have disastrous consequences. For instance imagine the following somewhat more straightforward proposal: the ASI should take those actions that would in expectation most fulfil the coherent extrapolated volition of all beings that would in expectation exist after its action. If this proposal were to be implemented, it would likely result in a reward hacking of the utility function. If the utility function of the ASI would be to maximize the fulfilment of the coherent extrapolated volition of all the sentient beings that would in expectation exist after its actions, then, the ASI could just create many digital minds with very simple extrapolated volitions to satisfy. If a sufficiently large number of these kinds of digital minds were created by the ASI's actions the force of their extrapolated volitions could be many orders of magnitude greater than the force of the extrapolated volitions of currently existing human and non-human animals in guiding the ASI's behaviour. This could result in the ASI disregarding sentient human and non-human animals and only performing those actions that would most satisfy the easy-to-satisfy preferences of the created digital minds. And these easy-to-satisfy preferences need not be preferences that seemed valuable to us (nor to our extrapolated volitions) in any respect. This alternative proposal of how to include future not-yet-existing sentient beings in the extrapolation base would likely lead to very undesirable outcomes. It is naive and flawed, because, at any given point in time it lets the ASI itself determine what new sentient beings come into existence (independently of the preferences of the extrapolated volitions of the already included individuals). Contrary to this, on the proposal I have put forward, what sentient beings come into existence, and shape the ASI's behaviour by being included once they come to exist, is entirely dependent on the preferences of the extrapolated volitions of the currently included individuals. Since the extrapolated volitions would know how SCEV works, they would be aware that in preferring a certain action or another they are also preferring the creation of some kinds of minds or others. Furthermore, they would also be aware that once they come to exist, they will be included in the extrapolation base and thus also affect the ASI's behaviour. Then, the extrapolated volitions would know better than to perform actions leading to scenarios similar to the one in which immense numbers of digital minds with invaluable and easy-to-satisfy preferences have been created and have complete control over the ASI's behaviour. Because of this, in the original proposal I uphold, the reward hacking is prevented.

This proposal would adequately take into account the interests of future sentient beings since, once they began to exist, their interests would directly be included in the ASI's utility function. Contrary to what seems to be the case in the standard CEV proposal, the interests of future not-yet-existing sentient beings, once they exist, would not be taken into account merely to the extent to which the extrapolated volitions of currently existing individuals desire to do so. And, by taking directly into account the interests of future sentient non-humans in this manner, the non-negligible chance of s-risks from the ASI's actions as a result of an adequate implementation of the standard CEV proposal would be significantly reduced.

The s-risks from the own ASI's actions are substantially lower when implementing SCEV because it is no longer the case that their prevention could only be caused by the extent to which humans who had their volitions coherently extrapolated cared about the suffering of the sentient non-human victims of the s-risks. Rather, the astronomical amount of volitions of the sentient non-humans themselves would directly be included in the extrapolation base. On this proposal, there still is (at least) a non-negligible chance that humans' extrapolated volitions would want to perform actions that would (without further intervention) lead to s-risks such as direct panspermia, or the use of simulation technologies. As a consequence, there is still some chance that these kinds of actions were actually performed by the ASI. However, once they were being performed, once new sentient beings such as non-human animals living in "natural" environments on other planets or digital minds came into existence, their extrapolated volitions would be directly included in the extrapolation base. And since their volitions would be extremely opposed to suffering immensely, the ASI would almost certainly prevent the astronomical suffering. Once the would-be victims of s-risks were to come into existence, their interests in not suffering would be directly taken into account and reflected in the ASI's behaviour. The astronomical suffering would almost certainly be prevented by the ASI and it would as with any other beings already present in the extrapolation base, try to fulfil the extrapolated volitions of the new existing beings. If the volitions of future sentient non-humans were included in the extrapolation base, as SCEV proposes, the occurrence of s-risks from the ASI's actions (i.e. what it causes or allows) would almost become impossible. Therefore, we have very strong pro-tanto moral reasons to implement SCEV to guide the behaviour of a possible ASI instead of only coherently extrapolating the volitions of currently existing humans (a minority of all the affectable moral patients expected to exist).

Finally, it should also be noted that this proposal of SCEV (as CEV) is not intended as a realist theory of morality, it is not a description of the metaphysical nature of what constitutes the 'good'. I am not proposing a metaethical theory but merely what would be the most morally desirable ambitious value learning proposal for an ASI. It is not the case that if moral realism is true, SCEV or CEV would necessarily arrive at this moral truth. Thus, (as CEV) SCEV and the arguments in favour of it are compatible both with a realist and an anti-realist conception of meta-ethics.

4. Objections

In this fourth section, I shall present and respond to three possible objections against the (pro-tanto) rejection of CEV and my proposal of Sentientist Coherent Extrapolated Volition.

4.1. The Risk from Predators' Violent Predispositions

Yudkowsky has contended that a reason against directly including sentient non-humans such as non-human animals in the extrapolation base is that it may result in consequences we do not like as a result of how different the non-human volitions are from ours (Yudkowsky, 2016). Then, it might be argued that it is not the case that SCEV is preferable to CEV since there is also the possibility that it might lead to very morally undesirable goals for the ASI to pursue, such as violent goals, since, for instance, many non-human animals living in the wild have goals and preferences in favour of predation. Given the very considerable degree of uncertainty we currently face regarding the exact nature of what would be CEV's extrapolation process in practice, we cannot completely rule out the possibility that this could have significant negative consequences. Then, it seems possible that the ASI could have weird, bad or violent goals as a result of some of the desires had by non-human animals living in the wild.

Even though it is true that we cannot be completely certain about this, I believe it is very plausible that this risk and its badness are very significantly lower than the s-risk from implementing the normal CEV proposal. When implementing CEV, the only possible preferences against sentient non-humans enduring astronomical suffering are those had by humans who to some extent care about the suffering of sentient non-humans. While, in implementing SCEV, all the volitions of the victims of the possible s-risks as a result of the violent preferences of predators, would count against the occurrence of such s-risks. By the mere fact of being sentient beings and being included in the extrapolation base, the volitions of the possible victims of such strange s-risks would primarily and almost entirely consist of the desire to prevent such suffering. Indeed as mentioned above, the desire to be free from immense amounts of intense and unbearable suffering is probably the strongest possible desire any sentient being can have, and thus since in an s-risk there would be astronomical amounts of beings having these desires, this would count extremely heavily against the occurrence of any kind of possible s-risks including the ones due to predatory dispositions.

Due to empirical uncertainty about what entities are sentient and black swans, it is likely that we can never be sure that the future occurrence of s-risks is completely impossible while sentient beings continue to exist and we expand through the universe or have sufficient computational power. But plausibly, one of the best ways to ensure that this does not happen would be to have an ASI that would truly take into consideration the interests of any of the victims of the possible s-risks. Demanding more certainty against the occurrence of s-risks than this is unwarranted, since it may indeed not be possible. Even if there is a very tiny probability that an s-risk could occur as a result of the predators' corrupted preferences, this probability is negligible. It would be unreasonable and too demanding to dismiss an ambitious value-learning proposal because it cannot with complete certainty rule out the possibility of the future occurrence of any kind of s-risk. And, since the probability and badness of the occurrence of bad outcomes as a result of predator's violent preferences is significantly lower than the (at least) non-negligible probability of the occurrence of s-risks as a result of implementing CEV, we would still have very strong reasons in favour of choosing SCEV if we could do so.

4.2. Democratic illegitimacy and being jerkish

Another possible counterargument to the general project of this paper of arguing for the importance of including the interests of sentient non-humans to guide the behaviour of a future ASI is that doing so would be anti-democratic. Most people indeed would likely not want nor vote in favour of implementing a value alignment proposal that equally takes into account the interests of non-human sentient beings.

This argument, however, assumes that only humans have claims in favour of having their interests represented in democratic procedures. This might be a result of speciesist bias or as a result of holding the previously mentioned view that only the interests of moral agents matter in a morally relevant way. However, all sentient beings could be positively or negatively affected by an ASI and accordingly, all of them have claims in favour of having their interests taken into account. It is unreasonable to believe, as we have seen, that those interests could be adequately represented only to the extent to which currently existing humans that are moral agents care about them. Thus, in this case, the principle of the all-affected interests in democratic theory clearly applies (Christiano & Sameer, 2022), and all sentient beings would have strong democratic claims in favour of having their interests directly represented.

This kind of objection is raised by Yudkowsky in a different manner. In his view, programmers or the implementors of a value learning proposal into an ASI should try not to be "jerks". And in doing so they should try to be fair and take into account the preferences of people that do not want all sentient beings to be included, by not directly including all sentient beings in the extrapolation base. Doing so, in Yudkowsky's view, would not be impartial since it would be rooting for the preferences of people who, for instance, care more about non-human animals (Yudkowsky, 2016). Contrary to this, however, Yudkowsky holds that it would, in fact, be jerkish to not include humans who, as sentient non-humans, are powerless in relation to determining what value alignment proposal is implemented. These humans possibly include children, existing people who've never heard about AI or people with severe physical or cognitive disabilities unable to act on and express their own views on the topic.

However, as seen above, it is not the case that there are no reasons to include sentient non-humans since they too can be positively or negatively affected in morally relevant ways by being included in the extrapolation base or not. The fact that many of the parties that have some power over which value learning proposal is implemented (i.e. some humans) do not care about these reasons does not mean that they hold no weight. It is not completely clear what Yudkowsky means by "jerkish" or "to be jerks", however, if it is to be understood colloquially, in the same manner, that it would be unfair and jerkish for powerless humans to be excluded, it is also unfair and jerkish for sentient non-humans to be excluded. Today, due to a certain amount of societal moral progress and expansion in the group of beings that society takes seriously into moral consideration, we tend to seriously include both powerful and powerless humans, and thus it seems intuitively jerkish to exclude them from the extrapolation base. But a few years ago it plausibly would not have felt this way. Imagine that the decision of which beings should be included in the extrapolation base had taken place many years ago. If they could do so, it would have been much more preferable that when considering what entities to include, these past decision-makers actually looked at the most plausible moral reasons in deciding what to do instead of doing that which was most socially and intuitively acceptable, which would have resulted in excluding many powerless humans. The same is preferable right now and in the future. If someone is able to decide what entities to include in the extrapolation base, they should look at the moral reasons in favour of including various kinds of entities even if they might not be the most intuitively or socially acceptable to include from the perspective of those in power. And as we have seen there are indeed strong reasons to include all sentient beings. A proposal that excludes sentient beings, then, is not impartial, but rather, would be a jerkish and unfair proposal.

Indeed, if Yudkowsky is right that it is important to exclude some sentient beings from the extrapolation base, this presumably is because the kinds and number of beings included in it can likely affect the ASI's behaviour. Then, this means that, in expectation, it is not the case that the interests of sentient non-humans would equally be taken into consideration in the ASI's behaviour independently of whether they are directly included in the extrapolation base or not.

Furthermore, Yudkowsky also contends that only including existing humans in the extrapolation base is more simple (Yudkowsky, 2016). While this is the case, it is not a reason in favour of the moral desirability of the proposal if we are able to successfully include all sentient beings. The fact that CEV is simpler than SCEV does not make it more morally desirable all else equal. Since doing nothing or only coherently extrapolating one volition is simpler than CEV but it is not more morally desirable.

Because of this, against the objection, it is not the case that pursuing this proposal would be anti-democratic, jerkish or unfair, but rather it would be a much more democratic, nicer and fair option. Implementing CEV instead of SCEV would not do justice to the direct democratic and moral claims of sentient non-humans.

4.3. Would SCEV preserve the identity of sentient beings?

It has been argued that the implementation of CEV would produce undesirable results because it would cause humans to cease to exist (Yampolskiy, 2022). Roman Yampolskiy has argued that by implementing CEV "[w]e would essentially agree to replace ourselves with an enhanced version of humanity as designed by AI", since, with the quick extrapolation jump involved in CEV, there would not be any continuity with our identity, and we would cease to exist (Yampolskiy, 2022). He presents the following argument: "we can say that current version of humanity is H_0, the extrapolation process will take it to $H_{10000000}$. A quick replacement of our values by values of $H_{10000000}$ would not be acceptable to H_0 and so necessitate actual replacement, or at least rewiring/modification of H_0 with $H_{10000000}$, meaning modern people will seize to exist". This argument can also be applied to criticise SCEV. It is indeed the case that some sorts of radical enhancements to the capabilities of humans, animals and digital minds can cause them to cease to exist by not being identity-preserving. However, radical enhancements to the individuals in the extrapolation base need not occur at all by implementing SCEV, and if they do (as it has been argued) they could occur in ways that are identity-preserving and aligned with the good of each of the individuals (Bostrom 2008: p.123–126; Paez & Magaña, 2023: p.23–25).

As Yampolskiy suggests, it is indeed the case that by performing volition-extrapolating upgrades to the models of the minds and volitions of the individuals in the extrapolation base, the resulting models will have different preferences and values from our own. However, this is in part what we are trying to accomplish when implementing a proposal such as CEV or SCEV, we are trying to prevent value lock-in (Yudkowsky, 2004; MacAskill, 2022). If we were to implement current values to direct the behaviour of an ASI, this would almost certainly constitute an immense moral tragedy, since all future generations would be stuck with antiquated values. It seems clear that it would have been a tragedy if past human civilizations from Ancient History or the Middle Ages locked in their values, and they persisted and were imposed for all of humanity's future by an ASI. Similarly, we should not believe that current values constitute the peak of moral progress and that there is no more progress to be made (Yudkowsky, 2004).

There would likely be a very significant distance between the values and preferences of the models of the coherently extrapolated volitions of sentient beings and the values and preferences of currently existing beings. However, since it prevents value-lock-in, the existence of this distance is preferable to its non-existence, and it would not cause currently existing beings to stop to exist. What would occur is that models of our minds and volitions would be created to guide the ASI's behaviour. These models would know more, reason better, and be more cooperative. As a result, they would necessarily have better justified, more coherent, and reasonable ethical views than the once we have. These better-informed, more thought-through and more justified ethical views would then guide the ASI's behaviour. It is implausible to believe, as this objection seems to imply, that what these better-informed, more thought-through and more justified ethical views would necessarily prefer is to kill all sentient beings and create new ones. One would only have reasons to believe that SCEV or CEV would recommend this if one already believes that killing all sentient beings or all humans and replacing them with new ones is what ought to be done. Disagreeing and discussing what ethical views or preferences would be had by the models is the same as discussing what ethical views are more reasonable, coherent, well-informed, etc. And the view that what ought to be done instead of any other physically possible option, is to kill all sentient beings and replace them with new ones is highly implausible and nowhere seriously defended. It is much more reasonable to believe that instead, the better-informed and more reasonable models could prefer for sentient beings not to be enhanced at all. Or could prefer that we could be enhanced in just, inclusive, equitable, autonomy-respecting and identity-preserving ways as it is prescribed by contemporary transhumanism (The Transhumanist Declaration, 2009; Savulescu & Bostrom, 2009). If this were done, the enhancements performed need not cause current sentient beings to cease to exist, since, as it has been argued, radical enhancements can be identity-preserving both for human and non-human animals if certain conditions are met (Paez & Magaña, 2023: p.23–25).

There are two major sets of theories about the kind of entities sentient individuals are and about their persistence conditions over time (Paez & Magaña, 2023: p.23–25). On the Somatic account, sentient beings are living organisms, and their persistence over time consists in maintaining the integrity of their metabolic processes (Olson, 1997; van Inwagen, 1990). On this view, there would be ways of performing, radical enhancements both onto human and non-human animals that would preserve their identity, one could enhance many of their capabilities without disrupting their metabolic processes. This account is about individuals with a biological substrate and thus does not apply to digital minds. But an analogous account, applied to digital minds could claim that maintaining the integrity of certain physical structural features of the digital substrate would be necessary and sufficient for them to persist over time. An account like this would also allow enhancements to be compatible with the persistence of digital minds over time. On the other major account, the Psychological Account, sentient beings are psychological entities who continue to exist only if they have psychological continuity over time (DeGrazia 2005; McMahan, 2002; Parfit, 1984). As it has been argued (Bostrom 2008:

p.123–126), this account is also compatible with enhancements being identity-preserving if the following conditions are met: the changes are in the form of adding new capacities or enhancement of old ones, without sacrificing preexisting capacities. They are implemented gradually over an extended period of time, and the new capacities do not prevent the preexisting capacities from being periodically exercised. Furthermore, the subject retains her old memories and many of her basic desires and dispositions and the subject retains many of her old personal relationships and social connections. And finally, in the case of humans and some sufficiently cognitively sophisticated digital minds each step of the transformation process is freely and competently chosen by the subject and the transformation fits into the life narrative and self-conception of the subject (Bostrom 2008: p.123–126). In the case of non-human animals and other digital minds who cannot freely choose and comprehend each step of the uplifting process, instead of these final conditions, “one may, alternatively, require the uplifting process not to undermine their control over their lives, irrespective of whether animals [or digital minds] can understand that possibility” (Paez & Magaña, 2023: p.23–25). This, then would not undermine their control over their existence but rather make it more robust. For a more developed discussion on why enhancements can be made compatible with non-human animals preserving their identity, see (Paez & Magaña, 2023: p.23–25) from where this paragraph is based.

Since we have good reasons to believe that radical enhancements to sentient beings can be identity-preserving, it is implausible to believe that instead of performing this kind of enhancements, if any, the more reasonable, coherent and well-informed ethical views had by the extrapolated models of sentient beings would prefer to kill all sentient beings by performing non-identity-preserving enhancement interventions instead. There are no good reasons to believe that this is preferable, and even if for some reason one were to believe so, it then would not constitute an objection to implementing SCEV, since, one would presumably prefer and welcome that which one believes to be preferable. The fact that SCEV would inscribe the values of $S_{10000000}$ (to represent all sentient beings instead of only humans) into the behaviour of the ASI, does not imply either the replacement or rewiring/modification of S_0 with $S_{10000000}$. Thus, we have no good reason to believe that SCEV would not preserve the identity of sentient beings.

5. Why it is unclear whether trying to implement Sententist Coherent Extrapolated Volition is best all things considered

I have argued that there are strong pro-tanto reasons to implement SCEV instead of CEV if we could do so successfully due to the fact that it would significantly reduce the non-negligible chance of s-risks from the own ASI’s behaviour that there is from an adequate implementation of CEV. However, in practice, it is not clear that these reasons are not overridden or defeated by other reasons against SCEV. In this section, I will lay out other possible pro-tanto reasons that go against trying to implement SCEV or a similar value learning proposal instead of implementing one which is closer to CEV.

First, there is the risk that the SCEV would not be implemented exactly as intended. When trying to implement SCEV there is always the possibility that we would not get everything right, and that, by accident, there could be unintended consequences. For some kinds of accidents from near misses that could occur, it seems plausible that the more morally desirable the ambitious value learning proposal, the worse the accidents that may result from trying to implement it. This is so because there are some possible kinds of near-miss accidents where the goals of the ASI identify those entities and things which can sustain value, but affect them in strange or bad ways contrary to desirable behaviour the ASI was intended to have as stipulated in the morally desirable ambitious value learning proposal. There are plausibly many possible ways in which these kinds of accidents could occur. One specific accident of this kind is if, through a value learning proposal such as SCEV, the interests of digital minds are taken into account and valued and as a consequence, many of them are created (as argued by Shulman & Bostrom, 2021), but we accidentally do not weight their suffering adequately. It is possible that even if the SCEV proposal intended to care about the suffering of digital minds, it accidentally was not able to adequately detect suffering and disvaluble states that many forms of alien digital minds might suffer (Vinding, 2018; Tomasik, 2019b). Or that it may not adequately weigh their interests by accident. This could potentially result in astronomical suffering. Such a kind of accident would be less likely if a less morally desirable ambitious value learning proposal were implemented since it would be less likely to create astronomical amounts of digital minds with the potential of suffering astronomically by accident. Another possible accident of this kind is the case of SignFlip, where the ambitious value learning proposal identifies a very morally desirable target or goal to maximize, but, by accident, the opposite target or goal is maximized (Tomasik, 2019b). In the case of SCEV, this would consist of maximizing the opposite of the coherent extrapolated volition of all sentient beings which would be extremely morally undesirable. Since there is some probability of the occurrence of these kinds of accidents, and some of the badness of the accidents would be mitigated by implementing a less morally desirable value learning proposal instead of SCEV, this gives us some other reasons against trying to implement a value learning proposal such as SCEV which comes very close or completely captures that which we value.

There are also further reasons against trying to implement SCEV in practice related to relevant game-theoretic considerations in specific decision-making situations. One possible and likely decision-making situation is one in which we ought to decide between implementing CEV or SCEV to a single ASI, but where the conditions for implementing SCEV instead of CEV are not fully optimal, where it is not the case that only one united decision-making entity is able to decide between the proposals, but rather, where there are already some decision makers that strongly prefer CEV instead of SCEV. In many of these kinds of decision-making situations, it may be much less desirable to try to implement a value learning proposal that adequately takes into account the interests of all sentient beings. Implementing SCEV may have even worse consequences in expectations for sentient non-humans due to a backlash from the opposition that may arise in trying to pursue this proposal. There is then, a strong reason against trying to implement SCEV due to the fact that, in practice, it might be net-negative in expectation.

However, it is not completely clear that this will be the case, and we might indeed find ourselves in future in which the pro-tanto reasons against trying to implement SCEV do not outweigh the pro-tanto reasons in favour of doing it. And, even if we can conceive of plausible future scenarios in which it seems to be the case that trying to pursue a proposal similar to CEV instead of SCEV would be preferable all things considered, to take that decision as a result of an adequate cost-benefit analysis of the different proposals it is crucial to understand the strong pro-tanto reasons in favour of SCEV that I have laid out in this paper.

6. Conclusion

In this paper, I have shown why we have some very strong pro-tanto reasons in favour of implementing SCEV instead of CEV. This is the case even if, all things considered, it is still ultimately unclear whether what is best is to try to implement SCEV or another proposal more similar to CEV. The action-relevant implications of what I have laid out in this paper, however, are not all contingent on the realization of the future in which we can fully decide what ambitious value learning proposals to try to implement. Even if we could be sure such a future will not be realized, there would still be some practical implications that follow. Determining how much more morally desirable SCEV is than CEV due to the strong pro-tanto reasons presented in the paper, is crucial to adequately make tradeoffs between different considerations in non-ideal decision-making situations where a value learning proposal has to be implemented. Furthermore, it is likely that the mere fact of having in mind what a possibly ideal scenario of adequate value specification would look like is useful in determining what we should realistically strive for if we could only reach more modest goals. Research into how different alignment and value learning proposals for possible future transformative AIs such as an AGI or ASI could affect sentient non-humans (who constitute the immense majority of present and future sentient beings expected to exist) is highly neglected. More research along these lines is required if we want to ensure that the far future goes well for all sentient beings.

Acknowledgements:

I am deeply grateful, for their suggestions, encouragement, and support to Olle Häggström, Anthony DiGiovanni, Magnus Vinding, Julia Persson, Lukas Gloor, Aluenda Diana Smeeton and an anonymous reviewer of this journal. This paper was conducted as part of an impact project at Future Academy, I am also very grateful to the organizers for having given me this opportunity. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Althaus D. & Gloor L. (2016). Reducing Risks of Astronomical Suffering: A Neglected Priority. *Center on Long-Term Risk.* https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/

Baum, S. D. (2020). Social Choice Ethics in Artificial Intelligence. *AI & Society*, 35(1): 165–176. DOI: 10.1007/s00146-017-0760-1.

Baumann, T. (2017a). S-risks: An introduction. *Center for Reducing Suffering*. https://centerforreducingsuffering.org/research/intro/

Beckstead, N. (2013). *On the Overwhelming Importance of Shaping the Far Future*. PhD, Rutgers University. https://rucore.libraries.rutgers.edu/rutgers-lib/40469/

Bostrom, N. (2008). Why I Want to be a Post Human When I Grow Up. In *Medical Enhancement and Posthumanity. The International Library of Ethics, Law and Technology*, edited by B. Gordijn and R. Chadwick, 2: 107–136. DOI: 10.1007/978-1-4020-8852-0_8

Bostrom, N. & Savulescu, J. (eds.) (2009). *Human Enhancement*. Oxford University Press.

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

Caviola, L., Everett, J. A. C., & Faber, N. S. (2018). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology*, 116(6): 1011–1029. DOI: 10.1037/pspp0000182

Chan, S. (2009). Should we enhance animals? *Journal of Medical Ethics*, 35 (11): 678–683. DOI: 10.1136/jme.2009.029512

Christiano, T. & Sameer B. (2022). Democracy. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/archives/spr2022/entries/democracy/

Christian, B. (2022). *The Alignment Problem: Machine Learning and Human Values*. W. W. Norton & Company

Clarke, S., Zohny, H. & Savulescu, J. (2021) *Rethinking Moral Status* (eds.). Oxford: Oxford University Press.

DeGrazia, D. (2005). *Human Identity and Bioethics*. New York: Cambridge University Press.

Dhont, K., Hodson, G., & Leite, A. C. (2016). Common ideological roots of speciesism and generalized ethnic prejudice: The social dominance human-animal relations model (SD-HARM): The social dominance human-animal relations model. *European Journal of Personality*, 30(6): 507–522. DOI: 10.1002/per.2069

Dvorsky, G. (2008). All Together Now: Developmental and ethical considerations for biologically uplifting nonhuman animals. *Journal of Evolution and Technology,* 18(1): 129–142. Available at: https://jetpress.org/v18/dvorsky.htm

Gabriel, I. (2020). Artificial Intelligence, Values, and Alignment. *Minds & Machines,* 30: 411–437. DOI: 10.1007/s11023-020-09539-2

Greaves, H. & MacAskill, W. (2021). The Case for Strong Longtermism. *Global Priorities Institute*. https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/

Groff, Z. & Ng, Y. (2019). Does suffering dominate enjoyment in the animal kingdom? An update to welfare biology. *Biology and Philosophy*, 34(40). DOI: 10.1007/s10539-019-9692-0

Harris, J. & Anthis, J.R. (2021) The Moral Consideration of Artificial Entities: A Literature Review. *Science and Engineering Ethics*, 27: 53. DOI: 10.1007/s11948-021-00331-8

Low, P. (2012). The Cambridge Declaration on Consciousness. *The Francis Crick Memorial Conference on Consciousness in Human and non-Human Animals*. Cambridge. https://fcmconference.org/img/CambridgeDeclarationOnConsciousness.pdf

McMahan, J. (2002). *The Ethics of Killing: Problems at the Margins of Life*. Oxford: Oxford University Press.

Ng, Y. (1995). Towards Welfare Biology: Evolutionary Economics of Animal Consciousness and Suffering. *Biology and Philosophy*, 10(3): 255–85. DOI: 10.1007/BF00852469

O'Brien, G.D. (2022). Directed Panspermia, Wild Animal Suffering, and the Ethics of World-Creation. *Journal of Applied Philosophy*, 39(1): 87–102. DOI: 10.1111/japp.12538

Olson, E. T. (1997). *The Human Animal. Personal Identity without Psychology*. Oxford: Oxford University Press.

Ord, T. (2020). *The Precipice*. Hachette Books.

Paez, E. & Magaña, P. (2023). A democratic argument for animal uplifting. *Inquiry: An Interdisciplinary Journal of Philosophy.* DOI: 10.1080/0020174X.2023.2248618

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

Proctor H. S., Carder G. & Cornish A. R. (2013). Searching for Animal Sentience: A Systematic Review of the Scientific Literature. *Animals (Basel)*, 3(3): 882–906. DOI: 10.3390/ani3030882

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Sebo, J. (2018). The Moral Problem of Other Minds. *The Harvard Review of Philosophy*, 25(1): 51–70. DOI: 10.5840/harvardreview20185913

Shulman, C. & Bostrom, N. (2021). Sharing the World with Digital Minds. In Clarke, S., Zohny, H. & Savulescu, J. (eds.) *Rethinking Moral Status*. Oxford: Oxford University Press.

Soares, N. (2016). The Value Learning Problem. *2nd International Workshop on AI and Ethics, AAAI-2016*. Phoenix, Arizona. https://intelligence.org/files/ValueLearningProblem.pdf

Sotala, K. (2012). Advantages of Artificial Intelligences, Uploads, and Digital Minds. *International Journal of Machine Consciousness* 4 (1): 275–291. DOI: 10.1142/S1793843012400161

Sotala, K. & Gloor, L. (2017). Superintelligence as a Cause or Cure for Risks of Astronomical Suffering. *Informatica* 41(2017): 389–400. https://www.informatica.si/index.php/informatica/article/view/1877

The Transhumanist Declaration (2009). *Humanity+*. Available at: https://www.humanityplus.org/the-transhumanist-declaration

The Universal Declaration on Animal Welfare, (2007). World Society for the Protection of Animals: https://www.worldanimalprotection.ca/sites/default/files/media/ca_-_en_files/case_for_a_udaw_tcm22-8305.pdf

Tomasik, B. (2015a). Artificial Intelligence and Its Implications for Future Suffering. *Center on Long-Term Risk.* https://longtermrisk.org/artificial-intelligence-and-its-implications-for-future-suffering

Tomasik, B. (2015b). Reducing Risks of Astronomical Suffering: A Neglected Priority. *Center on Long-Term Risk.* https://longtermrisk.org/risks-of-astronomical-future-suffering/#Some_scenarios_for_future_suffering

Tomasik, B. (2018). Will Space Colonization Multiply Wild-Animal Suffering? *Essays on Reducing Suffering*. https://reducing-suffering.org/will-space-colonization-multiply-wild-animal-suffering/

Tomasik, B. (2019a). What Are Suffering Subroutines? *Essays on Reducing Suffering*. http://reducing-suffering.org/whatare-suffering-subroutines/

Tomasik, B. (2019b). Astronomical suffering from slightly misaligned artificial intelligence *Essays on Reducing Suffering*. https://reducing-suffering.org/near-miss/

Tomasik, B. (2020). The Importance of Wild-Animal Suffering. *Center on Long-Term Risk.* https://longtermrisk.org/the-importance-of-wild-animal-suffering/

van Inwagen, P. (1990). *Material Beings*. Ithaca: Cornell University Press.

Vinding, M. (2018). Moral Circle Expansion Might Increase Future Suffering. https://magnusvinding.com/2018/09/04/moral-circle-expansion-might-increase-future-suffering/

Waldhorn, D. R. (2019a) Invertebrate sentience, summary of findings, part 1. Rethink Priorities. https://rethinkpriorities.org/publications/invertebrate-sentience-summary-of-findings-part-1

Waldhorn, D. R. (2019b) Invertebrate sentience, summary of findings, part 2. Rethink Priorities. https://rethinkpriorities.org/publications/invertebrate-sentience-summary-of-findings-part-2

Yampolskiy, R. V. (2022). On the Controllability of Artificial Intelligence: An Analysis of Limitations. *Journal of Cyber Security and Mobility*, 11(3): 321–404. DOI: 10.13052/jcsm2245-1439.1132

Yudkowsky, E. (2004). Coherent extrapolated volition. *Singularity Institute for Artificial Intelligence*. https://intelligence.org/files/CEV.pdf

Yudkowsky, E. (2008). Artificial intelligence as a positive and negative factor in global risk. In Bostrom, N. & Ćirković, M.M. (eds.) *Global Catastrophic Risks*. New York: Oxford University Press.

Yudkowsky, E. (2016). Coherent extrapolated volition (alignment target). Arbital: AI Alignment. https://arbital.com/p/cev/