Is artificial wellbeing a higher moral priority than future human wellbeing?

This essay argues that the answer to its title question is probably "yes". Let me define *AI sentience risk* as the following: The risk of a future in which (a) AI (artificial intelligence) systems have immense amounts of negative wellbeing (suffering) or (b) AI systems which could have had immense amounts of positive wellbeing do not come to exist. Since the importance of (a) is less sensitive to variations in ethical views, I will focus on (a) henceforth.

My main claim is this:

Thesis: Devoting resources to the reduction of AI sentience risk is better (in expectation) than aiming to improve the long-term future prospects of humans (and biological successor species), if one places high value on positive wellbeing or the absence of negative wellbeing.[1]

I cannot build a watertight case for *thesis*, since any conclusion about moral priorities is somewhat sensitive to one's beliefs regarding specific difficult moral and empirical questions.[2] However, I point to some general considerations which strongly support *thesis* and defend *thesis* against some objections. Given these considerations, *thesis* should be regarded as the default view which is only to be rejected if some ethical and empirical views by experts change surprisingly and to the detriment of *thesis*.

I present six arguments for *thesis*. These arguments presuppose that sentient AI is possible.[3]

1. The numbers argument: Envisioned future technology includes nano size AI, self-replicating AI and digital uploading of mental states. Thus, the number of potential future AI systems is plausibly multiple orders of magnitude higher than the number of potential future humans.

2. The utility monster argument: If sentient general AI is possible, then its design could plausibly be optimized to have the features which increase a being's capacity for wellbeing (perhaps, e.g., speed of conscious experience, pain intensity, varied emotions and so on).[4] Thus, the lowest wellbeing states of AI could be orders of magnitude worse than the lowest wellbeing states of humans.

[1] I take 'sentience' to be the capacity to have conscious experiences with a positive or negative valence, i.e., experiences which feel good or bad. Examples are (conscious) joy, relief, fear and pain.

[2] For instance, if one thinks that one (usually) ought to do what maximizes expected value or expected choice-worthiness (MacAskill et al. 2020), then the best course of action ultimately depends on specific degrees of beliefs and assignments of values to outcomes or choice-worthiness to options.

[3] To be more precise, the arguments presuppose that AI systems can have a wellbeing, i.e., that their existence can be good or bad for them. It is conceivable that sentience is not required for wellbeing.
Moreover, the relevant sense of 'possibility' is epistemic possibility, i.e., consistency with our knowledge. Whether sentient AI is possible in some objective sense is not decision-relevant since we don't have access to this information.

[4] For the notion of differences in wellbeing capacity, see Schukraft (2020) and Browning (2022).

3. The tractability argument: Mainly, the resolution of AI sentience risks depends on how humans choose to design AI systems, not on further facts about world history. Thus, people involved in the design of AI (scientists and regulators) may have much control over AI sentience risks. Moreover, there is a realistic change that sentience-related capacities and intelligence are sufficiently independent that a solution to AI sentience risk does not interfere with the development of powerful AI (which increases the compatibility of reducing AI sentience risk with political and commercial incentives).

4. The moral uncertainty argument: A scenario in which extreme numbers of AI systems (argument 1) suffer extremely (argument 2) would be very bad according to almost all axiological views discussed in the academic literature.[5] Rare exceptions would be (i) views which heavily discount future value and (ii) views which do not aggregate wellbeing between individuals at all *and* attribute equal moral weight to any being which fulfills conditions met by normal humans. Thus, *thesis* is very robust under axiological uncertainty.[6]

5. The comparison argument: The most influential argument for the reduction of existential risks has more force when applied to AI sentience risk. According to this standard argument, the future contains huge numbers of potential human lives. Thus, even a small chance to influence whether these futures lives come to exist and whether they are good has enormous expected value.[7] Since the number and wellbeing capacity of potential future AI is even higher, the standard argument supports *thesis* more strongly.

6. The bias argument: Many want to resist *thesis*. It feels implausible that our obligation to take steps to safeguard the wellbeing of AI might be higher than our obligation to ensure humanities flourishing. However, arguably this intuitive resistance is caused by the same anthropocentric biases which prime us to dismiss our obligations to animals, especially insects.[8] If so, we should discount our intuition that *thesis* cannot be true.

Let's combine these arguments: AI sentience risk has a massive scale (arguments 1 and 2), there is some reason to hope that it is tractable (3), its importance is robust to axiological uncertainty (4), the standard argument for the importance of existential risks applies to it with increased

[5] Axiology is the sub-discipline of philosophy that concerns value, i.e. good- and badness. The badness of suffering is usually regarded as self-evident. For a review of theories of fixed-population axiology, see Holtug (2015), and for a review of theories of variable-population axiology, see Greaves (2017).

[6] *Thesis* is not robust to uncertainty about theories of normative ethics, if the theories don't place a high value on improving total wellbeing (e.g., some deontological views). However, this sensitivity is shared by all arguments for moral views which are prominent in Effective Altruism.

[7] For examples of this basic argument, see Bostrom (2013) and Greaves and MacAskill (2021).

[8] For a discussion of these biases, see Mikhalevich and Powell (2020) and, briefly, Sebo (2021).

force (5) and there is a plausible explanation for why *thesis* seems nevertheless unappealing (6).

There are two objections to *thesis* which I reject: The first counterargument denies that AI sentience is possible. The second counterargument argues that – since there is no generally accepted theory of consciousness – we cannot know how to prevent AI suffering. The first objection exaggerates our certainty about AI consciousness, the second exaggerates our uncertainty. Since there is uncertainty about the correct theory of consciousness and most theories imply that some AI systems can be conscious,[9] there is a decent chance (at least 10%, but plausibly over 50%) that conscious AI is possible. Since we know that humans are sentient and we have some understanding of which states are conscious, we can make educated guesses about which AI systems are more likely to be suffering than others.[10]

The best objection to *thesis* is that the same course of action, namely ensuring that powerful future general AI (AGI) is aligned with human values, is the best way to mitigate both the risk of human extinction or suffering and AI sentience risk.[11] Arguably, whether a powerful AGI – if it arises – acts in accord with reasonable values is the main determinant of the course of future world history as a whole. For instance, powerful AGI might cause as well as prevent the creation of AI capable of suffering.

Two points mitigate this objection. First, the objection does not strictly target *thesis*, for it is compatible with the claim that the best use of resources aims to prevent AI sentience risk. It merely holds that the best way to address AI sentience risk is also effective to address risks for humans. Thus, this objection constrains the practical implications of *thesis*. Second, there remains nevertheless the risk that sentient AI arises before we have powerful AGI. Due to our profound uncertainty about both AI sentience and the timelines for the advent of AGI, this possibility ought to not be neglected.

Ultimately, it is an open empirical question how to best respond to AI sentience risk. Direct technical and advocacy work should very likely be prioritized more. The technical side involves work on what it takes for AI to be conscious and to suffer and how to build powerful AI which either is provably not conscious or has a reliably positive wellbeing. The political side might involve raising awareness among decision-makers about AI sentience risk and lobbying for

[9] The most influential contemporary theories of consciousness are all most naturally interpreted as entailing that an AI with the right kind of functional or physical organization would be conscious, even though it is not biological. These are: global-workspace theory (2020), integrated-information theory (2015), higher-order theory (2022) and recurrent-processing theory (Lamme 2018).

[10] See Shevlin (2020) and Dehaene et al. (2017) for attempts.

[11] A structurally analogous, but weaker, argument could be made in respect to attempts to improve institutional decision-making or to grow the effective altruism community. Both might decrease the risks of human extinction or suffering and AI sentience risk simultaneously.

extending legal protections to AI which can feel (similar to animal welfare law). Nevertheless, work on more general and indirect "risk factors", chiefly the alignment of AGI with good values, is very likely also part of the best resource allocation for a community (like Effective Altruism) which tries to maximize its marginal moral impact. Thus, the practical implications of accepting *thesis* for such a community, while profound, might be less revolutionary than they seem on first sight.

To condense my argument into two sentences: If sentient AI is possible, then its wellbeing mainly determines the expected goodness of the future. Our actions should reflect this fact.

Word count (excluding title, footnotes and references): 1199 words

References

Bostrom, N. (2013). Existential Risk Prevention as Global Priority: Existential Risk Prevention as Global Priority. *Global Policy*, *4*(1), 15–31. https://doi.org/10.1111/1758-5899.12002

Browning, H. (2022, January). The problem of interspecies welfare comparisons (preprint). Preprint. http://philsci-archive.pitt.edu/20115/. Accessed 4 May 2022

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science*, *358*(6362), 486–492. https://doi.org/10.1126/science.aan8871

Greaves, H. (2017). Population Axiology. *Philosophy Compass*, *12*(11), e12442. https://doi.org/10.1111/phc3.12442

Greaves, H., & MacAskill, W. (2021). The case for strong longtermism. https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf. Accessed 1 June 2022

Holtug, N. (2015). Theories of Value Aggregation. In I. Hirose & J. Olson (Eds.), *The Oxford Handbook of Value Theory* (pp. 267–284). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199959303.013.0015

Lamme, V. A. F. (2018). Challenges for theories of consciousness: seeing or knowing, the missing ingredient and how to deal with panpsychism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *373*(1755), 20170344. https://doi.org/10.1098/rstb.2017.0344

Lau, H. (2022). *In Consciousness we Trust: The Cognitive Neuroscience of Subjective Experience*. Oxford, New York: Oxford University Press.

MacAskill, W., Bykvist, K., & Ord, T. (2020). *Moral Uncertainty* (1st ed.). Oxford University Press. https://doi.org/10.1093/oso/9780198722274.001.0001

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Mikhalevich, I., & Powell, R. (2020). Minds without spines: Evolutionarily inclusive animal ethics. *Animal Sentience*, *5*(29). https://doi.org/10.51291/2377-7478.1527

Schukraft, J. (2020). Comparisons of capacity for welfare and moral status across species. *Rethink Priorities*. https://rethinkpriorities.org/publications/comparisons-of-capacity-for-welfare-and-moral-status-across-species. Accessed 30 November 2021

Sebo, J. (2021). *How to Count Animals, more or less*, by Shelly Kagan. *Mind*, *130*(518), 689–697. https://doi.org/10.1093/mind/fzz089

Shevlin, H. (2020). General intelligence: an ecumenical heuristic for artificial consciousness research? *Journal of Artificial Intelligence and Consciousness*. https://doi.org/10.17863/CAM.52059

Tononi, G., & Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1668). https://doi.org/10.1098/rstb.2014.0167