Why Reducing X-Risks Should Not Be a Major Concern of Effective Altruism[1]

Kwok Pak Nin, Samson

(Draft, please do not quote)

Abstract

In this paper, I argue that Effective Altruism should place less emphasis on reducing the risk of human extinction (reducing X-risks). My argument has two versions: ***a strong version and a moderate version***. The strong version adopts an anti-natalist stance, asserting that ***bringing new individuals into existence is ultimately unbeneficial*** because human life inherently involves significantly more bad than good. Additionally, if humanity were to cease existing, it would immediately ***remove the human-caused suffering of non-human animals and their descendants*** once and for all. Therefore, human extinction would be a positive outcome that EA has no compelling reason to prevent. In contrast, the moderate version does not hold such a pessimistic view of human life. Instead, it argues that it is more ***effective*** to focus resources on addressing current issues rather than uncertain future risks. Moreover, given the substantial uncertainty about whether future generations will have lives worth living, and the pressing crises we face today, the ***moral cost*** of allocating significant resources to uncertain future risks—rather than tackling immediate problems—would be too high for EAs to justify. In conclusion, reducing X-risks should not be a major concern of Effective Altruism.

[1] I would like to express my sincere gratitude to Dr. Chong Yun Chak, Dr. Wong Kim Fan, and Mr. Loi Man Ho for their valuable comments and insights.

Introduction

According to Peter Singer, Effective Altruism (EA) focuses on several areas to do good, such as reducing extreme poverty, alleviating the suffering of non-human animals, and minimizing the risk of human extinction (reducing X-risks). As a consequentialist, I strongly support the fundamental principles of EA and its efforts to combat extreme poverty and reduce animal suffering. However, I argue that EA should place less emphasis on reducing X-risks[2], since ***it is not the best way that we can do good***.

EA's Position on Human Extinction

First and foremost, I would like to clarify the position of EA regarding the risk of human extinction. EA seeks to make it a better world for both current and future beings. From this perspective, ***having new, happy individuals in the world is seen as a positive outcome that enhances overall well-being***. This stance can be interpreted in two ways:

1. ***If*** future generations exist, we should ensure that they will be living well, thus increasing total well-being.
2. We should ensure that the ***future generations, whose lives will be worth living, can come into existence***, thus increasing total well-being.

It is reasonable to assert that no rational individual would reject the first interpretation. However, it is important to examine the rationale for the second interpretation, which can be further divided into two forms regarding ***whose well-being*** is being promoted:

2a. We must ensure that future generations exist because their existence would enhance ***our*** total well-being.

2b. We must ensure that future generations exist because ***their lives*** would, on average, be worth living, thus contributing to total well-being.

Regarding point 2a, there are self-interested reasons to be concerned about the existence of future generations. As Scheffler (2018) observes, "we have self-interested reasons to ensure humanity's survival. We need future generations to survive in order

[2] In this paper, I will focus mainly on die-off extinction, which refers to population declines that occur without intentional killing, such as those caused by natural disasters or a lack of reproduction. In contrast, kill-off extinction involves the deliberate killing of individuals, leading to a premature end to humanity. While my arguments primarily apply to die-off extinction, I will leave the issue of kill-off extinction open for the reader's consideration.

for us to lead what I call 'value-laden lives'—lives structured by wholehearted engagement in valuable activities." However, I will not repeat these arguments here.

In contrast, point 2b presents a more ***altruistic motivation***. I interpret Professor Singer as supporting this perspective, especially when he references Derek Parfit's claim that the future of our species will be "extremely long and, on average, well worth living." Singer has also expressed that, as a hedonistic utilitarian with a cautiously optimistic view of moral progress, he sees the extinction of Earth-originating intelligent life as a significant loss. Furthermore, it appears that many members of the EA community share this viewpoint and have invested substantial resources in reducing X-risks. Therefore, a detailed examination of this interpretation is necessary.

The central question is: ***Why should we assume that future generations will be "extremely long and, on average, well worth living"?*** EA is fundamentally optimistic about our ability to alleviate suffering through targeted interventions. Historical advancements in healthcare, education, and social policies have consistently improved quality of life. EA seeks to accelerate this progress, suggesting that the lives of future individuals could be even more fulfilling. Therefore, if humanity were to face extinction, the ***potential realization of this optimistic world would be eliminated***, which would constitute a significant loss. However, I believe this perspective is problematic. I will outline my objections in two parts.

My Objection: The Strong Claim

The strong claim adopts an anti-natalist position and challenges EA's efforts to reduce X-risk by arguing that human extinction would be a ***positive outcome*** for future generations. Those familiar with anti-natalist views are likely aware of David Benatar's Quality of Life (QoL) argument (Benatar, 2006, 2013, 2015), which contends that ***all human life involves significantly more suffering than happiness***, making it morally wrong to bring new individuals into existence. Here, I will provide a simplified version of the QoL argument to illustrate my point.

While EA is correct in asserting that advancements in healthcare, education, and social policies will continue to extend human life and reduce physical suffering, several factors may still hinder this progress:

- Q1. ***Inherent Mental Suffering***: Human existence is intrinsically accompanied by mental suffering arising from negative emotions such as anxiety, depression, and interpersonal conflicts. These forms of suffering are deeply embedded in the human psyche and are not fully eliminable through technological advancements, education, or social policies. Even under optimal external conditions, individuals experience unfulfilled desires, fear of death, and

existential angst. Since these negative emotions are fundamental aspects of human consciousness and self-awareness, they persist regardless of improvements in material well-being. Therefore, mental suffering is an inherent and significant component of human life.

- Q2. ***Asymmetry Between Pleasure and Pain***: There is an empirical asymmetry in how humans experience pleasure and pain, where negative experiences outweigh positive ones both in intensity and duration. Pleasurable moments are typically fleeting and require considerable effort to attain and maintain. In contrast, painful experiences—such as illness, injury, loss, aging, and death—are often prolonged and occur without any effort. Moreover, many negative experiences are unavoidable aspects of life, whereas positive experiences often necessitate continuous striving and are not guaranteed. This asymmetry suggests that over a lifetime, individuals endure more suffering than they experience pleasure due to human life's vulnerability.

It is important to note that the two points discussed above pertain not merely to empirical circumstances of some people but to the ***inherent nature of human life***. These considerations apply universally to any context, as long as one is a member of the species *Homo sapiens*. Furthermore, if EA responds by claiming that we can aspire to a future where advancements transform humanity into a state resembling a god-like existence characterized solely by pleasure and free from suffering—despite my scepticism regarding this possibility—such a condition would ***no longer*** represent human life. In this sense, humanity would have already experienced extinction.[3]

If we accept the QoL argument, we must also recognize that human lives are much worse than many believe, and that all lives involve significant suffering. As a result, procreation should be viewed as a ***substantial harm*** to future generations. EA correctly points out that the future holds immense potential due to the duration and sheer number of individuals who will exist. However, it is crucial to acknowledge that this long-term future could be ***either*** extremely valuable or extremely detrimental, depending on the quality of life experienced by future individuals. Since I believe both Q1 and Q2 are valid claims, if EA genuinely cares about their well-being, they should not prioritize the reduction of X-risks.

Moreover, the strong claim extends beyond human concerns to include ***non-human animals***. Given that EA is committed to the welfare of all sentient beings, it is crucial to recognize that humans are the primary agents responsible for a significant

[3] I recognize that some proponents of effective altruism may welcome such a post-human future, as life in that context could be significantly more worthwhile. However, whether a life that consists solely of pleasure, without any suffering, constitutes a fulfilling existence is a complex question that I cannot address within the scope of this paper.

portion of animal suffering. Each year, billions of animals are bred, confined, and slaughtered for food, with many enduring severe sufferings in factory farms while still alive. Additionally, humans inflict considerable harm on animals through experimentation, entertainment, sport hunting, and habitat destruction. While efforts to reduce these harms are commendable, the complete elimination of such practices is improbable in the foreseeable future. If humanity were to cease existing, the infliction of suffering upon animals by humans would end immediately. This would not only alleviate the current suffering of countless animals but also prevent future generations of animals from experiencing human-caused harm. In essence, without humans, the suffering that animals endure at our hands would not occur. Arguably, ***human extinction represents a genuine form of animal liberation***, as it would permanently remove the primary source of deliberate animal suffering from the planet.

To conclude, the strong claim insists that reducing X-risks ***would not benefit*** future generations (both human and animal), so it should not be the task of EA.

My Objection: The Moderate Claim

The moderate claim recognizes the possibility that future generations may, on average, lead lives worth living. Therefore, reducing X-risks ***could be beneficial*** for them. However, this claim does not render the position of 2b plausible for EA. The rationale for this lies in the need to evaluate both (1) the ***probability*** of this counterfactual scenario and (2) the ***timeframe*** required for its realization.

While I acknowledge that reducing existential risks could, in theory, have a high expected value due to the vast number of future lives at stake, I believe that ensuring future generations, on average, lead lives worth living in our imperfect world is a challenging endeavor with a ***low probability of success***, particularly in the near term. Achieving this goal would require significant resources and radical changes across physiological, technological, cultural, and political spheres. Moreover, the probability of successfully mitigating existential threats—such as unaligned artificial intelligence, nuclear war, or global pandemics—is highly uncertain, as these risks are complex and influenced by ***myriad unpredictable factors***. In contrast, near-term interventions have high probabilities of success and deliver immediate, tangible benefits. For example, The Global Fund reports that its investments have saved 65 million lives from AIDS, tuberculosis, and malaria since 2002 (The Global Fund, 2024). These interventions have proven track records, and their outcomes are relatively certain. Allocating resources to these areas ensures a high expected value due to the combination of a ***high probability of success and significant immediate impact***.

Another aspect of the moderate claim insists that the ***moral cost*** of allocating substantial resources toward reducing X-risks instead of addressing the suffering of currently living individuals would be too high for EA to justify. While we aim to create a future where most sentient beings can live fulfilling lives, pursuing this goal may come at the expense of alleviating the suffering of many people today. For instance, every dollar spent on preventing potential risks could instead be used to help those who are suffering from poverty and disease right now. ***The longer we postpone addressing immediate suffering in the quest for an ideal world, the more we allow current suffering to persist***. Therefore, we face a critical choice: prioritizing actual suffering versus potential risks. The expected value of addressing present suffering is significant enough to warrant reconsideration.

To sum up, the moderate claim insists that even if reducing X-risks would benefit future generations, it is ***not the most effective way to do good***; therefore, it should not be a major task of EA.

Conclusion

There are many complex empirical calculations and counterarguments to my position that I cannot address in this context. In particular, the moderate claims require strong empirical support and statistical evidence to be decisive. However, my primary aim here is to raise important questions for us all to reconsider how we can do good. I believe that if we move beyond an anthropocentric perspective, the value of reducing X-risks may be ***less significant than it initially appears***—not only for EA, but for all of us as well.

References

Benatar, David. (2006). *Better Never to Have Been: The Harm of Coming into Existence*. Oxford University Press.

Benatar, David. (2013). "Still Better Never to Have Been: A Reply to (More of) My Critics." *Journal of Ethics*, 17(1–2), 121–151.

Benatar, David, & Wasserman, David. (2015). *Debating Procreation: Is It Wrong to Reproduce?* Oxford University Press.

Scheffler, Samuel. (2018). *Why Worry About Future Generations?* Oxford University Press.

The Global Fund. https://www.theglobalfund.org/en/results/