Guidelines for Discussing Longtermism and Existential Risks

Written by Nick Beckstead with input and feedback from various community members and several EA organizations. These guidelines are endorsed by the following organizations and individuals:

80,000 Hours, CEA, CFAR, MIRI, Open Phil, Nick Bostrom, Will MacAskill, Toby Ord, Carl Shulman

The guidelines were written primarily with the organizations in the effective altruism community writing about longtermism and existential risks in mind. They focus on making our discussion of existential risks and longtermism more attentive to the interests and concerns of the suffering-focused ethics community. EAF has written its own set of guidelines intended for people writing about longtermism from a suffering-focused perspective.

This document was originally shared with staff members at the above organizations and a handful of other individuals writing on these topics. The document is intended to be shared with individuals where it seems useful to do so, but was not intended for publication. To share the document with additional people, please request permission by emailing Bastian Stern (bastian@openphilanthropy.org).

Last updated: July 3, 2019.

Our recommendations

When communicating about the long-term future, we recommend that you go out of your way to:

- Highlight beliefs and priorities you have that are important to the SFE community, even if they may be of secondary importance to you. For example:
    - Acknowledge the possibility that the long-term future might turn out to be highly suboptimal - perhaps even containing significant tragedy/suffering - conditional on avoiding premature extinction.
        - Potentially flawed futures might result from: legacies of suboptimal cultures, some AI alignment failures, conflict involving AGI systems, or singletons led by suboptimal principles.
    - Consider referring to non-extinction-focused strategies for increasing the probability of high-value long-term futures. Potential examples include: solving the alignment problem, preventing AI arms races, and trying to create a more inclusive and reflective culture.
    - Avoid making simplifications that might leave your readers confused on these points. If you aren't deliberately arguing otherwise, try to avoid leaving your

reader with the impression that “longtermism” is only about reducing extinction risk, or that the future will inevitably be near-optimal and free from astronomical harms if extinction is avoided.

- Highlight assumptions that are required for your claims to be true that the SFE community might disagree with.
    - Example assumptions might include: (i) assigning high value to the existence of many flourishing future generations (rather than no such future generations existing), (ii) holding that future lives have positive average expected welfare overall.
- **Where appropriate, highlight ideas for mitigating astronomical harms that might occur.** For example, you might discuss approaches to AI alignment that are optimized for reducing harm, or whatever else you think could be good for reducing such risks.
    - This is recommended only when: (i) you find it comfortable and plausibly helpful, (ii) you are talking to audiences that might be receptive (such as advanced audiences that are especially open to unusual ideas), and (iii) you follow the SFE guidelines.
    - I’ve put this in bold because it’s of special importance to EAF.
- Highlight concepts that push against forms of extremism, such as moral uncertainty, moral trade, and the unilateralist’s curse.
    - As noted in the SFE guidelines, naive forms of consequentialism and unilateralist thinking can result in actions that cause direct real-world harm, or that cause reputational fallout for the effective altruism community that drastically reduces our ability to do good in the future.
    - Discussing matters of cosmic and world-historical importance in the vicinity of consequentialism raises these risks. Consider addressing this risk by referring to articles on moral uncertainty, moral trade, the unilateralist’s curse, and consequentialist case for good behavior. Potential references include:
        - Bostrom (2009): Moral uncertainty – towards a solution?, EA Concepts: Moral uncertainty, MacAskill (2014): Normative uncertainty
        - Bostrom (2013): The Unilateralist’s Curse. The Case for a Principle of Conformity
        - Christiano (2016): Integrity for consequentialists
        - Ord (2015): Moral Trade
        - Tomasik (2014): Reasons to Be Nice to Other Value Systems
        - Yudkowsky (2008): Ends Don't Justify Means (Among Humans)
- Familiarize yourself with the SFE perspectives on longtermism so that you can think about how to follow the above principles. Readings recommended by EAF include:
    - Reducing Risks of Astronomical Suffering: A Neglected Priority
    - S-risks: Why they are the worst existential risks, and how to prevent them
    - Cause prioritization for downside-focused value systems
    - Superintelligence as a Cause or Cure for Risks of Astronomical Suffering
- If you are writing a foundational work on longtermism that you think might be referred to extensively in the future, consider asking for feedback from someone who is familiar with the s-risk perspective (such as Jonas Vollmer or Lukas Gloor). EAF reports that they’d “be happy to give feedback on relevant writing from this

perspective whenever it's useful." They are also available to comment on documents about AI macrostrategy if you think that would be useful.

We intend these guidelines to be both truth-promoting and cooperative. Where there are conflicts between cooperation and promoting the truth that you can't resolve,[1] we would like you to side with promoting the truth. We would consider it counterproductive if trying to follow the above recommendations led you to:

- Treat ideas as more plausible than you actually think they are.
- Privilege hypotheses you deem insufficiently plausible to discuss.
- Cite work you don't think deserves to be cited on its merits.
- Voice support for proposals that you find unpromising.

Firm stance against advocating wrongful aggression

There are a variety of actions that almost everyone would regard as harmful, wrong, and/or dishonorable. We don't know exactly how to characterize them, or exactly what the philosophical basis for their wrongness is. Actions in this category include (but are not limited to) certain forms of deception, promise-breaking, manipulation, harassment, and violence. Almost everyone knows these actions in these forms are wrong, and most moral philosophers try to explain this in ethical theories (despite pushback from some act consequentialists). Let's call the principles that forbid these things, "common sense ethics."

There is room for healthy debate about what is and is not part of common sense ethics. For example, is a soldier who voluntarily kills enemy combatants in a plainly unjust war in violation of common sense ethics? Is someone who argues in favor of a pre-emptive war arguing for violating common sense ethics? These cases may not be straightforward, but many cases are straightforward, and we will handle them straightforwardly.

We wish to draw a bright line around straightforward and severe violations of common sense ethics that involve violence. It might be argued that using violence in a way that violates common-sense ethics could, in one circumstance or another, be part of an optimal plan for reducing existential risk/s-risk, helping the global poor, or protecting animals. Such claims generally seem highly dubious to us. Whatever the merits of these claims, violent actions of this type - especially when done in the name of doing as much good as possible - are simply unacceptable to the effective altruism community. Perpetrators or advocates of such actions are not welcome in our community. If you are aware of individuals advocating or pursuing such behavior, we strongly encourage you to report it to Claire Zabel or Julia Wise.

If you become aware of it, we'd also appreciate it if you let Claire or Julia know when someone is seriously considering or seriously discussing actions of this type. While we aren't

[1] Apparent conflicts between following these guidelines and promoting the truth may partly be addressed by being clear about why you are discussing the issues that you have gone out of your way to discuss.

advocating that such consideration/discussion be grounds for removal from our community, we do want to be aware of it early in case it develops further.

This position mirrors a stand EAF is taking.

Clarifying examples

In order to illustrate the notion of "straightforward and severe violations of common sense ethics that involve violence," I work through some clarifying examples, non-examples, and borderline cases below. One simple test I am using to determine whether a violent action is a violation of common sense ethics is to ask whether I expect it to be illegal.

- Straightforward violation: Someone robs a bank at gunpoint in order to donate money to a noble cause. (Widely regarded as wrong, severe, involves violence, illegal.)
- Straightforward non-violation: Someone steals a laptop from a table while the owner is in the bathroom, sells it, and uses the money for a noble cause. (Widely regarded as wrong, illegal, and arguably severe, but not covered because not violent.)
- Straightforward non-violation: An executive director of a charity lies to donors in order to help their organization succeed. (Widely regarded as wrong, arguably severe, and possibly illegal, but not covered because not violent.)
- Straightforward non-violation: Person A shoves person B at EA Global in order to get them to stop interrupting their conversation. (Widely regarded as wrong, violent, and possibly illegal, but not covered because not severe.)
- Straightforward non-violation: A researcher does painful experiments on mice in an attempt to find new drugs for chronic pain conditions. (Arguably violent and arguably severe, but not covered because it is not widely regarded as wrong; also not illegal.)
- Not straightforward, so not covered: An animal welfare activist breaks into a factory farm to rescue hens and document their conditions. While there, she is attacked by a guard. She severely injures the guard while defending herself. (Violent, severe, and illegal; unclear whether it would be widely regarded as wrong because it involves self-defense, and the details may turn on the circumstances of the case.)

To be clear, many of the "non-violations" above are wrong and should be heavily discouraged by the EA community, they just aren't the focus of the present document.