# EAF's Guidelines for Discussing Astronomical Stakes (shared)

*Written by the Effective Altruism Foundation (EAF) and the Foundational Research Institute (Jonas Vollmer, Stefan Torges, Lukas Gloor, and David Althaus) with input and feedback from various community members and several EA organizations.*

*First written: March 2019; last updated: 10 September 2019.*

*Please ask Stefan (stefan.torges@ea-foundation.org) before sharing this document further; he'll usually be able to get back within two days. Thanks!*

## Summary


- We've developed communication guidelines for effective altruists focused on reducing suffering in the long-term future. Following these guidelines will help others with different values a lot, thereby leading to large gains from cooperation.
- In particular, we would like to avoid (inadvertently) contributing to strong forms of future pessimism because this could inspire unilateralist acts of aggression and violence.
- We encourage you to adhere to the following guidelines in public communications:
    1. Minimize the risk of readers coming away contemplating causing extinction, i.e., discuss practical ways to reduce s-risks instead of saying how the future could be bad,
    2. Go out of your way to be charitable and balanced,
    3. promote cooperation and compromise, and
    4. contextualize vivid descriptions of actual or potential future suffering.
- We do not tolerate severe violations of common sense ethics that involve (violent) aggression in the name of reducing suffering and will take action against this.
- The case for doing these things rests on (1) preventing harm according to other people's values, (2) preventing harm according to the values of our own community, and (3) engaging in concrete positive-sum collaborations with others in the effective altruism community.


## Introduction

As a community of people dedicated to reducing suffering, we have always taken great pains to act and write in a manner that reduces the chance that unilateral actors will seek to disruptively and actively increase extinction risks or engage in other violent behavior, which we expect would have adverse consequences for our community and the world at large. We consider this the responsibility of anybody who writes about these topics.

We believe that cooperative communication as outlined in this document is one of the most cost-effective efforts of helping effective altruists focused on a flourishing future, and we think it will contribute to trust and large gains from cooperation. We believe that makes the extra effort easily worth it.

In response to us writing these guidelines, Nick Beckstead wrote guidelines for EAs focused on reducing existential risks to encourage other EA organizations (80k, CEA, CFAR, FHI, GPI, MIRI, and Open Phil) to communicate about the long-term future in a way that's more cooperative towards suffering-focused EAs. In general, other long-termist EA organizations have been very cooperative and are taking concrete steps to support work on s-risks more.

While we think the benefits from cooperation are large, we think many of the recommendations listed here would be worth following even without such cooperation, as doing so also reduces the risk of reputational damage, increases trust in relevant networks, and helps the SFE community attract talent and funding.
With these guidelines we don't presume to speak on behalf of the whole SFE community. Rather, we see our role as trying to facilitate cooperation with others. Your actions contribute to the extent of the reciprocation that will be part of this cooperative effort. So we hope that, as a community, we can coordinate around these guidelines and pull in the same direction.

# The Case for Cooperative Communication

## Goal: Minimizing the number of people holding the NNT belief combination

People holding the following set of beliefs (**NNT**) are particularly likely to engage in harmful behavior:

- **(1) Negative Future.** Given the current trajectory of human civilization, the future will likely contain more disvalue than value.
- **(2) Naive consequentialist unilateralism.** One should always maximize expected value from one's own narrow perspective—disregarding common sense heuristics and even going against the views and values of the majority of other people and effective altruists.
- **(3) Taking ideas seriously.** It is good to take one's explicit beliefs and goals literally and seek to creatively and unconventionally optimize for them.

Holding the NNT belief combination doesn't necessarily entail harmful action (this also depends on other beliefs), but it greatly increases the risk of harmful action.

We already discourage (2) through our writings and actions. (3) is a core idea of effective altruism that we don't want to discourage in general. However, as a direct or indirect result of our work, people might come to believe (1). Since some people might come to hold (2) despite our best efforts, we might indirectly counter the work of other effective altruists by contributing to a wider acceptance of (1). We can avoid inadvertently increasing the number of actors who hold this set of beliefs by reducing advocacy for (1) and discouraging (2).

## Promoting the NNT belief combination would harm others

Those holding the NNT belief combination might attempt to (directly or indirectly) impede efforts by effective altruists who are interested in as positive a future as possible. For that reason, we think promoting (parts of) the NNT belief combination would be uncooperative.[1]

In general, the case for moral cooperation is very strong and convincing, as it ranges from commonsensical heuristics to theory-backed principles found in Kantian morality or throughout Parfit's work. Reasons for it can also be found in the literature on decision theory and be derived from multiverse-wide superrationality. Even determined consequentialists should strive to be cooperative and trustworthy.

Common-sense morality strongly prohibits all acts of illegitimate violence or sabotage as they violate rights and cause direct harm. Effective altruists and society in general are deeply opposed to such acts.

In addition, acts of aggression would likely hurt the entire effective altruism community, not just those working towards positive futures.

## Promoting NNT would directly harm our own cause

It is very likely that attempts of bringing about extinction or otherwise curtailing the upside potential of human civilization would fail or backfire. This would plausibly increase the amount of suffering as opposed to reducing it.

Concretely, we would likely be excluded from the most valuable networks and work on AI alignment because of mistrust people have formed about our community or what they perceive to be our community. We would also suffer reputational damage, both within the general public, but also the effective altruism community, making it much harder to recruit exceptional people, attract large donors, and be effective in general. If this led to us being a despised group, any future work would be impossible.

## Not promoting the NNT belief combination is positive-sum

As the main group of people advocating for reducing suffering in the EA community, EAF is in a particularly good position to encourage good communication norms. Doing so will build trust and enable positive-sum collaborations with other effective altruism groups. Specifically, we expect this will enable regular joint research retreats with top AI alignment researchers, increased access to large donors and the best talent of the community, and advice from senior researchers.

In a similar effort to increase cooperation and compromise, we gave input on content written by other effective altruists organizations to ensure that suffering-focused concerns are

[1] Note that we're not taking a stance on whether moral advocacy in general is (un)cooperative; our primary concern is harmful unilateral action by people holding the NNT belief combination.

adequately reflected in writings from people who don't share these values themselves. We are confident that others are interested in communicating cooperatively and with a positive-sum mindset.

# How to Talk about Astronomical Stakes

We encourage you to follow these guidelines for all forms of public communication, including personal blogs, social media, essays, books, talks, meetups, and scholarly publications. When in doubt, what matters is the spirit of the guidelines, not the letter.

## 1. Minimize the likelihood of people contemplating causing extinction

Whenever we communicate about suffering-focused ethics or s-risks, we think we have the responsibility to ask ourselves: Could somebody come away from this thinking that it might be a good idea to cause human extinction or disruptively interfere with efforts to prevent it? Also consider telephone game effects: How might other people misunderstand or misrepresent your ideas, and could this cause them to reach such conclusions? As outlined above, this is a real risk, and everyone writing about these topics can contribute to reducing it, often at low cost.

Content might be particularly risky if efforts to increase extinction risk or to undermine extinction risk prevention are framed as morally praiseworthy, the option of destroying the world is made salient, or specific methods for achieving it are suggested. If these ideas are at the core of the text, we'd encourage you to reconsider publication. If they might be inferred from the text, we'd encourage you to consider ways of mitigating the chance of people contemplating causing extinction.

In general, we recommend writing about practical ways to reduce s-risk without mentioning how the future could be bad overall. We believe this will likely have similar positive results with fewer downsides because there are already many articles on theoretical questions.

You can also circulate such content as google docs or PDF files among researchers without uploading them to the public internet. This preserves many of the benefits without the risks. EAF has successfully managed to raise awareness of s-risks in this way.

## 2. Discuss our views in a balanced way

If you're set on discussing more fundamental questions despite the arguments and concerns outlined above, we encourage you to go beyond academic norms to include some of the best arguments against these positions, and, if appropriate, mention the wide acceptance of these arguments in the effective altruism community. Many reasonable people in effective altruism and academic philosophy disagree with our normative views. Below we provide some examples of the relevant literature in this context, but we trust your judgment to

reference the most appropriate sources (e.g., some of these might not be appropriate for the academic context):

- For **normative** questions, you could consider referencing Beckstead: On the Overwhelming Importance of Shaping the Far Future, Pummer: The Worseness of Nonexistence, or Shulman: Moments of Bliss for alternative views. Consider emphasizing normative uncertainty (or the anti-realist equivalent of valuing further reflection), e.g. by referencing Bostrom (2009): Moral uncertainty – towards a solution?, EA Concepts: Moral uncertainty, MacAskill (2014): Normative uncertainty, Greaves & Ord: Moral uncertainty about population axiology. In the context of near-term suffering, it may be more cooperative (and also more convincing) to adopt a "quality of life" or "everyone agrees suffering is bad" framing, rather than explicitly referring to suffering-focused views.
- For **empirical** questions, you could consider referencing Paul Christiano: Why might the future be good? or Ben West: An Argument for Why the Future May Be Good, and Shulman: Are pain and pleasure equally energy-efficient?. Note that most researchers focused on reducing suffering believe the long-term future will be very positive on empirical grounds, such that the best case for focusing on reducing s-risks stems from fairly strong normative assumptions or neglectedness considerations.
- **We do not ask you to lie or distort your views.** You can reference opposing views and note their wide acceptance and plausibility without necessarily endorsing them. Whenever there are conflicts between cooperation and honesty that you can't resolve, we would like you to side with honesty.

There are many positive examples of this kind of balanced, cooperative, and careful writing, e.g., the section *Moral uncertainty and cooperation* in Gloor: Cause prioritization for downside-focused value systems.

Relatedly, we're in the process of developing a new definition of s-risks more focused on the most important scenarios from our perspective, i.e., on the outcomes with the highest expected badness (similar to how the definition of x-risks refers to a drastic curtailing of humanity's potential, not simply any small decrease in the expected value of the future). We expect to change the definition by the end of 2019. If you're planning to write about s-risks, please get in touch so we can coordinate.

## 3. Promote cooperation & compromise

We think promoting cooperation and compromise is a good practice and helps decrease polarization, increase trust, and facilitate collaboration. We welcome further work on cooperation and moral trade, linking to existing resources, and emphasizing these ideas in future writings where appropriate.

Many people in the community have already written about the topic from different perspectives, e.g.:

- Yudkowsky (2008): Ends Don't Justify Means (Among Humans)
- Bostrom (2013): The Unilateralist's Curse. The Case for a Principle of Conformity

- Tomasik (2013): Gains from Trade through Compromise
- Tomasik (2014): Reasons to Be Nice to Other Value Systems
- Ord (2015): Moral Trade
- Christiano (2016): Integrity for consequentialists
- Schubert (2017): Moral cooperation
- Oesterheld (2017): Multiverse-wide Cooperation via Correlated Decision Making
- Todd (2018): Doing good together — how to coordinate effectively, and avoid single-player thinking

Consider acknowledging that there are other ways of positively shaping on the long-term future beside reducing s-risks: reducing extinction risks and ensuring a flourishing future. Doing so will encourage others to write more about practical ways to reduce s-risks.

## 4. Contextualize vivid descriptions of suffering

Graphic descriptions of actual or potential future suffering might inadvertently increase the likelihood of rash emotional decisions leading to harmful consequences, and contribute to a distorted view of how likely specific bad outcomes are. So while they may be important for emphasizing how bad certain negative experiences can be, we'd still encourage you to only use such descriptions where they're integral to the core point you want to make. When you do decide to include them, consider adding appropriate contextualization and caveats to minimize the risk of overreaction and distortion.
The same applies to vivid descriptions of dystopian futures.

## Checklist

We've condensed these points into a brief checklist. While they might be obvious for some people, we hope this list will still serve as a useful resource.

### Before writing/preparing

- **Think about your topic**. Could you capture many of the upsides, but incur significantly fewer downsides by discussing practical ways to reduce s-risks without making more salient how the future could be negative?
- **Weigh risks and benefits**. Is the added benefit of your new publication worth the added risks, especially in light of existing resources that may already contain similar points?
- **Consider engaging in adversarial collaboration**.

### Before publication

- **Model your least sophisticated reader**. Could they come away thinking that the future is more negative than you yourself think, or that increasing the risk of extinction might be the right course of action?
- **Check how balanced your writing is**. Have you stated and referenced alternative viewpoints charitably?

- **Consider asking for feedback from people who hold views that are different from your own.**
- **Consider soliciting feedback from Jonas Vollmer or Lukas Gloor**. They will always be happy to help.

# Direct Advocacy for Wrongful Aggression is Unacceptable

While we are not aware of specific instances of wrongful aggression, we still want to take an unambiguous stance against this form of behavior.

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.[2] It might be argued that using violence in a way that violates

[2] **Clarifying examples:**
In order to illustrate the notion of "straightforward and severe violations of common sense ethics that involve violence," Nick Beckstead has worked through some clarifying examples, non-examples, and borderline cases below. One simple test to determine whether a violent action is a violation of common sense ethics is to ask whether one can 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

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 Jonas Vollmer. We will investigate the issue and take appropriate action. If you become aware of it, we'd also appreciate it if you let Jonas know when someone is seriously considering or seriously discussing actions of this type. While we aren't 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. We will do our best to strike the delicate balance between facilitating open discourse and discouraging the unacceptable instigation of aggression in violation of common sense ethics. Reasonable individuals who approach us in confidentiality will have their confidentiality respected.

# Conclusion

Your contribution is important: Putting these guidelines into practice will lead to more productive collaborations in the EA community and allow us to reap the gains from cooperation. Please let us know if you have thoughts on (parts of) these guidelines. We want to make sure that we can all pull in the same direction instead of pursuing counterproductive strategies. Thanks a lot for taking the time to engage with these guidelines!

---

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.