Thresholds for funding existential risk interventions

Nuño Sempere

Edited by Spencer R. Ericson on October 16, 2024

Abstract

I try to add some practical clarity to the setting of a medium-sized altruistic funder seeking to spend a limited budget on interventions to reduce existential risk, where the value of interventions is hard to estimate. First, I discuss different estimation strategies for speculative areas, and I settle on sanity checks and thresholds as good initial steps. I then review some currently existing thresholds, and find that they are seeking to answer two distinct questions: 1) what amount of preventing existential risk would beat more well-trodden global development interventions?, and 2) how much should an altruistic funder in fact be willing to pay, given a limited budget? The two questions could have different answers if the marginal existential risk intervention beats the marginal global development intervention by a large margin. I then propose some thresholds in the second sense. Finally, I discuss some downsides of thresholds as a tool.

Contents

0. Editor's note

SoGive commissioned the following piece of work from Nuño Sempere. We intend for this work to not only help our clients, but to spark a more rigorous, quantitative discussion in the existential risk reduction community on cost-effectiveness.

SoGive broadly endorses Nuño Sempere's conclusions in this piece. However, it remains primarily the work of Nuño. SoGive has remaining critical uncertainties about some pieces of the models in this collection of pieces, such as the share of AI-based existential risk attributable to the UK and the share of risk reduction attributable to CSER.

SoGive readily invites debate, feedback, and forks of the model. Reasonable readers may disagree with several of the parameters used in these models. The public version of the GitHub repository for this work will be linked in the comments. This way, commenters can make their own versions of the model with their own parameters.

We have tried to make this project as open-source as we can, only removing personally identifiable interviews and rough drafts from the working repository to the public repository. In this way, we hope to encourage other organizations in the community to publish their models, so we can work together to converge upon the best estimates of existential risk and cost-effectiveness.

SoGive would like to thank Rethink Priorities for peer-reviewing this piece. We would like to thank Sam Hilton for edits. We would like to thank the Centre for the Study of Existential Risk for their comments, time, and cooperation. These parties may disagree substantially on some of the conclusions in Nuño's report as commissioned by SoGive. Mistakes remain attributable to the author, Nuño, and primary editor, Spencer R. Ericson.

1. Review of estimation strategies

1.1. Summary table

Strategy	Feasible?	Recommended?
1. Randomized trials	Impossible	No
2. Shapley values	Even more impossible	No
3. Natural experiments	Unlikely	No
4. Direct reduction in x-risk	With effort	Maybe
5. Relative value comparison	Yes	Maybe

Strategy	Feasible?	Recommended?
6. Estimate intermediate outputs	Yes	Yes
7. Sanity checks or thresholds	Yes	Yes
8. Other funders as a benchmark	Yes	Not really
9. Expert intuition	Yes	In moderation
10. Impact rubric	Yes	No

1.2. Discussion of strategies

1.2.1. Ideal yet impracticable

Randomized trials are the gold standard in medicine and global health and development. However, they do have the problem of double-counting or triple-counting impact. **Shapley values** address that limitation.

However, randomized trials are impossible[1] for the case of existential risk[2]; we can't boot up different possible worlds and try different existential risk reduction strategies in each. Shapley values are even more impossible, because we'd have to boot up many different counterfactual worlds.

Still, randomized trials and Shapley values could represent an aspiration: they are what other methods will seek to approximate. In particular, to very roughly approximate Shapley values, we can remember to adjust estimates of impact downwards in proportion to the number of people needed, or the number of steps in the pathway to impact.

One first step away from the ideal randomized trial would be to look at proxies and intermediate outputs, rather than at whether a catastrophe was averted. For example, instead of running an impossible randomized trial on whether existential risk is averted, one could run a less impossible randomized trial looking at an intermediate outcome, like whether laws are passed that seek to reduce existential risk. There is some sense in which that is less impossible, because one could look at lobbying across different times and places. But in practice, this method doesn't work either, because the vast majority of existential risk interventions have a small sample size. Still, for some interventions, like fellowships, one could try to aggregate multiple cohorts.

What's more, not many existential risk organizations have carried out randomization to estimate their impact. Still, we could hope for **natural experiments**. For instance, for a given fellowship, we could look at fellows right above and right below their bar for acceptance, and see if those who were barely accepted did better or worse than those who didn't.

[1] At best, we could do randomized trials over proxies and intermediate outcomes, e.g., if a fellowship accepts individuals, we could run an RCT over a few of them, perhaps repeated over a few cohorts to get a larger sample size.

[2] Throughout this document, we will be considering existential risk for conceptual simplicity and clarity. However, many of the same points apply to existential risk, however defined, which conceptually also includes e.g., stable totalitarian dictatorships.

1.2.2. Methods at the sweet spot of feasibility, informativeness, ambitiousness

Directly estimating how much catastrophic risk or existential risk a given intervention avoids is probably going to be too complicated. Still, one could look at the pathway to impact of the intervention, estimate their chances of success at each step, and estimate the magnitude of success if they do succeed.

In doing so, you will end up leaving the confines of scientific practice and enter the region of subjective Bayesian estimation, where you are quantifying your subjective beliefs of impact, but you might not convince others that your chosen quantification makes sense. Push even further, and you yourself might not trust the estimates you have produced.

Still, the quantification effort could be interesting, and you could learn something along the way. And direct existential risk estimation seems enticing: even though it will be very speculative and uncertain, it would provide a straightforward point of comparison between different interventions.

We could also ask several people about how much they **value different interventions relative to each other**, as in here or here. From these estimates, we can construct a linear scale and then place the different interventions in it to represent how much we value them relative to each other. This will capture participants' intuitions of value. But is time-intensive, and it can be opaque without even more effort

The academic literature has explored similar methods, under the keyword "discrete choice theory". However, they typically ask about binary choices, not about estimates of distributional relative value. Their objective is sometimes to estimate the value of features, e.g., of a renovated bathroom or bigger windows in the market for houses in a given city.

Beyond direct existential risk estimation and relative values, we could **make estimates of intermediate outputs**, perhaps using midway units, like quality-adjusted:

- …research papers
- …research insights
- …policy engagement
- …mentoring
- etc.

Each of these is a proxy for the ultimate impact. And we'd hope that these proxies scale roughly linearly. Or that if some organization has better proxies than another organization, they will be better overall.

But when comparing what quantity of one unit is better than what quantity of another unit, we are back to having to have some sort of relative value or direct existential risk comparison.

We could remember that ultimately, we want to make some decision, for example, whether to fund some project or not, and we could directly look for **sanity checks, or thresholds which would make our decision clear**:

- Is the project strictly worse or strictly better than some other project that we could fund instead?
- Is the organization not doing anything at all?
- Is the organization absurdly cost-inefficient?
- Is the organization absurdly cost-effective?

Particularly for the later two questions, it would help to have a threshold above which some existential risk intervention should *definitely* be funded, and a threshold below which it should *definitely not* be funded.

1.2.3. More common status quo methods

Although the above methods are sometimes used, it's more common to use lesser methods. There is a temptation to not do quantified estimates, and instead settle for **impact rubrics**. By this I mean aggregating some observable metrics around the world using some sort of checklist or Likert scale. Outside of longtermism, Animal Charity Evaluators and Charity Entrepreneurship do this. The problem with this approach is that values aren't linear: an intervention that scores twice as high on the checklist isn't therefore twice as valuable. And it might even be worse, if the checklist doesn't weigh more valuable factors comparatively heavily.

More often, grant evaluators simply rely on **expert intuition**, which relies on an implicit bar for funding. This could work if experts have good judgment. In practice, there isn't enough estimation infrastructure that one can move away from relying on the judgment of individuals soon.

Finally, there is the temptation of using **other funders as a benchmark**, for instance the Long-term Future Fund. This makes sense in that, if you can't do better than a fund, it makes sense to donate to it. But on the other hand, other funders may have idiosyncratic constraints or different beliefs, and taking other funders as a threshold could lead to game of chicken dynamics[3] or recursion problems[4]. In the case of the LTFF in particular, they tend to spend around an hour per grant, so it's not clear that it's worth deferring to them.

[3]With two funders and three interventions, such that the first funder values A > B > C but the second values A > C > B, if the first funder waits, the second funder funds A first, and the first funder can fund C.

[4]If different funders are deferring to each other, you could have weird dynamics where they aren't funding an intervention because maybe other funders have information that make them not fund a given project. For what it's worth, my impression is that this sometimes happens in the context of venture capitalists funding startups.

2. Review of past thresholds

2.1. List of past thresholds

- A question on the Effective Altruism Forum asked "How many EA 2021 $s would you trade off against a 0.01% chance of existential catastrophe?".
  - Linch's answer: ~$100M to ~$1B per basis point, based on gut feeling on the pool of grants.
  - Khorton's answer, based on short-termist thresholds: $3.5B, based on the value of saving lives of current living people
  - My own answer from back then, also based on short-termist thresholds: Upwards of $70B.
- Thomas Kwa has an unpublished Google doc, outlining his estimates of everything in EA. To derive those prices, he backtracks from Linch's estimate. The Google doc is private, but its comments contain a fair share of discussion on this topic.
- This 80,000 Hours interview with Ajeya Cotra of Open Philanthropy mentions that OP estimated a value of $200 trillion per world saved, or $20B per basis point.
  - As an editorial note, note that it seems that this $200 trillion estimate is in the context of setting a threshold which would beat more well-trodden global development philanthropy, not in the context of what threshold one should set given a limited budget to reduce x-risks. That is, one could imagine thinking that spending $20B to reduce existential risk by one basis point would beat spending $20B on global health development, while still thinking that one can do much better than averting one basis point of existential risk for $20B.
- Vasco Grilo collects some other individual estimates here. Spencer Ericson has a similar but slightly non-overlapping list here.
- Vasco also gathers some Rethink Priorities' estimates from their cross-cause cost-effectiveness model here. In particular, their default parameters involve a 0.6% reduction in existential risk for $8B to $20B, or 1 basis point for $1.6M to $4M (plus some chance of a negative outcome). However, this isn't presented as a threshold but rather as an illustration of how to compare hypothetical existential risk interventions with hypothetical global health interventions.

2.2. Meaning of these thresholds

There are various ways you can put a dollar value to existential risk prevention:

1. by giving a dollar value to DALYs saved. The question you are answering is: how much value (in DALYs, or in "dollars") would saving the world save? The answer looks like: amount of people who would live * dollar value of life / existential risk averted. The "dollar value of life" here will be something pretty high, e.g., 50K-100K per year, or even higher, because a lives are valuable, something we cherish: if I were *forced* to choose between

100K and an additional year of my life, and if I didn't have any cheaper options, I don't know which one I'd go with.

2. through comparison with near-termist thresholds. The question you are answering is: if your willingness to pay is the same as for near-termist interventions, how much should you be willing to pay for existential risk interventions? The answer here looks like: amount of people who would live * about $5K per life / existential risk averted. Here, the value of a life represents the cheapest you can save a counterfactual life. This question might help you prioritize between longtermist and near-termist interventions.
3. through thinking about what your willingness to pay should be. The question here is: if I deploy my dollars as far as they can go to reduce existential risk, what is the cost effectiveness? The answer looks like dollars spent at the best interventions your (marginal) dollars can buy per unit of existential risk averted. Here, what matters is how cheap you can buy the cheapest basis point of existential risk prevention.

For deciding whether to donate to more well-trodden global health and development interventions or to more speculative existential risk interventions, the relevant question to be asking is #2.

However, when deciding to which existential risk intervention you should donate, the relevant question is #3.

I suspect that there has been substantial confusion on this topic, leading to inflated valuations of longtermist projects. As an analogy, I at some point estimated that good headphones provided me with 6K to 100K of value, in terms of better mental health and increased productivity (this is high, but they are absurdly effective for me in both cases). If I had no other option, I'm pretty sure I would pay $6k for good headphones. However, I *do* have other options, because you can get some excellent headphones for less than $500. Similarly, if you "value" an expected future life at $5k, the value of speculative existential risk interventions would seem to be very high. But that is not the correct calculation to be making for *willingness to pay*. For willingness to pay, you want to be spending your whole budget paying as low a fraction of a penny per future life as possible[5].

3. A few tentative thresholds

3.1. From first principles

3.1.1. A minimum willingness to pay Is there some amount such that one should be willing to pay at least that amount to avert a given amount of existential risk? I say yes. We can build such a minimum willingness to pay bound by noticing that it would be a good outcome if the existential risk

[5]This is complicated by money not necessarily being the limiting reagent.

community spent all its money[6] to bring existential risk[7] down to zero. If we do so, we arrive at a threshold of ~$4M (1.49M to 13.2M) per basis point (0.01%) of existential risk reduced.

This threshold should be understood as an ambitious amount: if you can find an opportunity that buys 1 basis point for ~$4M, my guess is that that funders seeking to reduce existential risk should be eager to take it.

The two basic building blocks of the model are estimates of total funding and total risk. As the capital seeking to reduce existential risk rises or the amount of risk decreases, it can afford to relax its cost-effectiveness while still meeting its raison d'être. As one increases capital, at some point one would reach a point where one would want to solve existential risk and then have extra funding left over. Then this threshold would no longer be valid.

3.1.2. An maximum willingness to pay Humanity's ability to coordinate to avert catastrophes is bounded. Its ability to coordinate would be highest for an unambiguous, verifiable threat, like an asteroid heading directly towards Earth, or an unmistakable alien invasion which declared war beforehand. Maybe in those scenarios, humanity would be able to mobilize the whole world, for instance under a greatly expanded United Nations for multiple decades. In that sense, we could say that humanity's willingness to pay to avoid existential could be some multiple of the world's current GDP ($~100T/year). For example, we could say that humanity's willingness to pay to avoid existential was $1,000T if humanity was able to coordinate for ten years to spend all of its manpower deflecting an asteroid.

In practice, humanity would not be able to bring that level of coordination to threats like powerful and misaligned AI, pandemics, or nuclear war. In that sense, the maximum amount that humanity could spend on those types of threats is much lower than many multiples of the world's GDP.

So an upper bound of one's willingness to pay is given by the largest check that one could conceivably cash. Consider a "mega Manhattan project" that mobilized 10 to 100 million people, paid each wages of $20k to $200k for 5 to 50 years, and was modeled to bring risk down from 100% to 0% risk. This seems like the best humanity could do in the worst scenario. This would correspond to a maximum willingness to pay for a basis point of ~$3B (400M to 23B).

However, we've arrived at that amount considering civilizational willingness to pay. But it's unclear if that can directly translate to willingness to pay for a much smaller donor. Strictly speaking, that upper bound on civilizational

[6] Here, I estimate it as $3B to $15B. The lower bound is considering that Open Philanthropy has around $10B, of which not all will be allocated to existential risk prevention. The upper bound considers the possible existence of additional funders.

[7] Here subjectively estimated as 7% to 30% over the vaguely defined medium term (e.g., the next 100 years). To replicate this, one could try to aggregate previous estimates, e.g. this database of existential risk estimates.

willingness to pay doesn't logically imply an upper bound for much smaller donors. For example, it could be that a much smaller community is able to coordinate more tightly and make different trade-offs, for instance because it has much lower discount rates and so values the future much higher.

However, while that could be possible, in practice getting less than a basis point of reduction for 400M to 23B means that the solutions wouldn't be able to scale to tackle the existential risk being addressed even if the whole of humanity put its efforts into it. This seems like a loud warning sign of gross inefficiency, and hence the upper threshold stands. More generally, I just expect reasonable existential risk interventions to exceed this threshold, and so I think it is useful to quickly discard potential interventions.

3.2. Thresholds based on better specific interventions

The previous thresholds were based on abstract reasoning. We can complement that abstract reasoning with comparisons with existing projects. In particular, if we find some existing, scalable project whose returns don't diminish or diminish slowly with scale, then we can check that a proposed project meets that bar.

This might be akin to how GiveWell compares proposed top interventions against GiveDirectly. However, the situation is not analogous, since existential risk interventions are by and large much more speculative than GiveDirectly. And so we may not want to endorse potential points of comparison as much as GiveWell endorses GiveDirectly, and not imply that returns diminish analogously slowly as for GiveDirectly.

3.2.1. Why not the LTFF

Initially, I was keen on using the LTFF as a reference point.

However, in practice I found out that this wasn't a great idea:

- The LTFF does have reasonably steep diminishing returns, so it's a moving target and not ideal for a threshold.
- The LTFF is very focused on risk from AI, and a large fraction of its funding is going towards funding technical AI safety. But then it feels more meaningful to model the impact of marginal funding to AI technical safety—for instance through MATS—and not bother with modeling the LTFF as an intermediary.
- The LTFF spends very little time per grantee, and it seems like an organization like SoGive, which can spend much more time per grantee, should be able to find more cost-effective grants.

In addition, I hold some idiosyncratic beliefs which reinforce this decision:

- I disagree with the LTFF's emphasis on technical AI research. I have a higher level of baseline skepticism, implicitly expect future shock to be unlikely or muted, and find myself more worried about unknown unknowns.

- Based on some unpublished research, I think that the LTFF doesn't have a great track record of success. The reader can perhaps get some intuitions by browsing through the largest grants.
- The LTFF has a variable value. Some years, its bets will pay off. Other years, they will fail to do so. And this is variable enough that, in arbitrary, relative units, LTFF in 2022 could have a value of 0.1 and LTFF in 2024 could have a value of 10. But this makes it unwieldy as a threshold.

3.2.2. Comparison against robust AI safety technical research

Consider funding a few talented PhD students to do technical AI safety research, for instance as in MATS. This intervention seems reasonably scalable and robust[8], so we might want to use it as a threshold as well. That is, if a proposed intervention isn't as good as technical AI safety research, we might as well fund technical AI safety research instead.

One way to model the value of such research is by thinking about what the positive value of the AI safety community has been, estimate the total size of that community, and then estimate the value of new entrants as a proportion. Here, a problem is that maybe the AI safety community has had negative value. However, because AI technical safety research is relatively harmless, I think it makes sense to consider the positive component.

With this in mind, if we take:

- existential risk reduced by the AI safety community[9]: 0.01% to 2%
- size of the AI safety community[10]: 300 to 10k.
- cost per person: \$70k to \$150k/year
- Average career duration so far: 3 to 7 years

Then we arrive at a cost per basis point of between \$1.8M and \$1.25B.

Now, I like this model because thinking in terms of the impact of the whole AI safety community, their total impact will be large enough that I can intuitively grasp it. In contrast, if I were directly modeling the impact of one individual researcher, I would have to multiply the chance that they are working in an agenda that has some chance of being conceptually correct or adequate, and then some chance of counterfactually avoiding existential or catastrophic risk. That later number would be pretty small and difficult to conceptualize. I prefer the top down in this case.

[8] In the sense that it likely doesn't increase existential risk, not in the sense that its value is certain.

[9] Where is this estimate coming from? Well, risk reduction is not completely insignificant (hence above 0.01%). But it's definitely lower than 50%, 20%, 10%, 5%. I estimated 2% as a reasonable upper bound.

[10] An estimate from 80,000 hours from 2022 puts the number of full time equivalent people working in AI safety as ranging from 100 to 800. That estimate has the upside of going organization by organization. But it is from 2022, and since then, interest has risen. Additionally, you will have research that is applicable to AI safety even if it doesn't go by that name, hence the upper bound of 10k people.

We can then take MATS[11] in particular as proof that an intervention that fits the above desiderata exists—though there will be more of such interventions, e.g., restricted funding to the LTFF for technical AI safety, another new AI safety fund, etc. MATS in particular, due to its selection effects, technical focus, strong mentorship program, etc., is probably averting existential risk at a cost-effectiveness point equal or higher than the cost-effectiveness of the AI safety community on average. This is important because we might think that most of the impact of the AI safety community comes from highly selected outliers, and so it's important that MATS does seem to contain outlier technical talent as well.

One problem with using the AI safety community as a threshold, though, is that although I model it as so above, it's not clear that its value is positive. For instance, through its existence it could bring to the attention of amoral actors that AI is likely to be *powerful*, and those actors might seek to invest in AGI development, maybe because at some level they value their private benefit over the collective existential externality. Consider that the rationality community was involved in the creation of each of DeepMind, OpenAI and Anthropic, and that their successes have attracted more amoral mimics, like Mistral, Inflection AI or Stable Diffusion.

To address the possibility of that downside, one could:

- restrict potential grants to AI safety grants which one takes to be robustly good. For instance, within technical safety, interpretability seems like it has low potential for harm. However, that doesn't address the specific pathway mentioned above of attention to safety signaling potential power to others.
- take the hit in terms of expected value and accept that there is some chance that a grant could be negative
- consider that the ship of not bringing more attention to AI has already sailed, and that even if bringing attention to AI through increased attention to AI safety was negative in the past, it is not so now.

A final difficulty is that this threshold changes with the amount of funding available. For instance, another billionaire could come in and fully fund all plausible AI technical safety research. If so, this threshold would have to be updated or discarded.

3.2.3. Other interesting targets for comparison

- ALLFED. One might hope that they might be conceptually easier to model than AI. However, modeling their specific kind of R&D is difficult because their impact in the event of a crisis depends on a) being heard and b) someone else putting their ideas into action.

[11] Previously SERI MATS, but the program has since become independent of the Stanford Existential Risk Initiative.

- GovAI might be amenable to modeling as a result of its soothingly well-outlined theory of impact.

3.3. Summary table and link[12] to the models

Threshold	Meaning	Decision criteria
Minimum willingness to pay threshold	Ambitious threshold, beating this is a very good sign	If an intervention beats this, fund enthusiastically
Robust AI technical safety work threshold	If you don't meet this threshold, there is a reasonably scalable intervention you could donate to that will have better results	If an intervention beats this, fund. This depends on donor preferences, and could be tweaked
Maximum willingness to pay threshold	Very weak threshold, not beating this is a sign you might be Pascal mugged or grossly inefficient.	If an intervention doesn't beat this, definitely don't fund. Useful to discard interventions early on on research.

3.4. A note on replicability of estimates

In a recent revision, I've tried to give some more specifics about how I'm arriving at my estimates. In the interest of replicability, here are some notes of possible approaches to take next time SoGive is trying to do some of these estimates:

- Look at the specific models. In these reports, I mention results and hint at methods, but the models, and in particular the files ending in .c, have the actual numbers, and often comments pointing to sources.
- Look at the world to get a very rough lower and upper bound. Then, fit a suitable distribution (often, lognormal for unbounded distributions and beta distributions for probabilities and quantities bounded between 0 and 1).
  - Because of the magic of Fermi estimates, many wide distributions multiplied together will often produce narrower distributions, so getting a narrow distribution here is often unimportant.
- If you want some more intersubjectivity, you could consider polling people you respect. If so, you might want to structure that polling as Delphi estimate, where participants first write down their own estimates and give reasons, then review others' reasons, then update their own estimates.
- Often, someone has already published estimates for the quantity you are interested in. In particular, the EA Forum archives are searchable.

[12] Note that for now this is a private Github repository, and so isn't accessible to the public. This should be changed when this project is completed.

- Consider delegating. You could hire forecasters, or people interested in estimation, like Vasco Grilo, Samotsvety, or myself, to produce estimates for you.