Evaluation of CSER (WIP)

Nuño Sempere

February 13, 2024

Abstract

I present positive and negative aspects of CSER, and then present a synthesis: CSER as a tricky opportunity for impact. I resolve that trickiness by modelling its impact, and applying a threshold model. I discuss modelling choices and caveats, and conclude with recommendations.

Contents

1. Overview of CSER

1.1. Introduction and positive aspects of CSER

The Centre for the Study of Existential Risk, CSER, is a 23 person strong think tank. It sees itself as doing strong work around existential risk, but also as myth-busting and critical questioning of the rest of the existential risk community. Perhaps as result, they not only work on existential risk proper (extinction and scenarios similarly bad, e.g., stable global dictatorships), but also work on catastrophic risks (scenarios which could cripple but not kill humanity) and "merely" large catastrophes.

Their primary pathway to impact, at a high level is, I think, as follows:

- Produce research on existential risk
  - To directly gain insights
  - To grow and nourish an academic research field on the topic
  - To increase the seniority of its researchers and affiliates
  - To gain credibility from policy-makers
  - To act as a critical interlocutor, and as a bridge between more mainstream worries, like climate change, and more speculative worries, like risk from advanced AI. Correct the existential risk community when it is wrong or overeager.
- Interact with policy-makers, and provide advice
  - Do this as part of a whole existential risk community, because many voices have more credibility than only one voice
  - Organize workshops, recently also send participants to secondments
  - Be more left-wing than right-wing; it's possible this could pay off in a future Labour government[1].
  - Affiliation with Cambridge is useful for credibility for policy-makers
  - Presence in the UK is a bet on the UK's relevance in AI. The AI summit validates that bet so far.

Per this interview with Seán Ó hÉigeartaigh[2], policy impact for the AI team seems to be particularly high.

1.2. Neutral aspects of CSER

CSER has some "marshmallow test" aspects. It looks expensive, but that expense pays. For instance, having more than one think tank advocating similar

[1] At some point during my investigation, I was more pessimistic about CSER as a whole, and then thought that this point could end up being more important. This is because some level of CSER being suboptimal could be outweighed by some level of it being more influential in some future labour government.

[2] Seán Ó hÉigeartaigh is a mouthful, so I'll just refer to him as Seán.

policies increases robustness, which has paid off after the successive FHI scandals. Affiliation with Cambridge leads to increased cost, but also to closeness to policy impact. My sense is also that CSER might be in a better position, politically, after the next UK elections, which Labour is positioned to win.

CSER can also be seen as building bridges between, on the one hand, more speculative existential risks (existential risk from runaway artificial intelligence or from pandemics, etc.) and on the other hand, more "mainstream" existential risks (climate change) and more left-wing preoccupations (algorithmic bias). Given that CSER's work on the latter two areas is funded by other funders, like the Templeton or Grantham foundations, this could be an opportunity for synthesis and cross-pollination, while remaining cost-effective.

CSER has occasionally offered critiques of and to the broader existential risk community. For instance, some researchers at CSER became convinced that democratization was key to improving decision quality in existential risk. They didn't convince the EA or existential risk communities, leading to somewhat of an impasse. On the other hand, if their critique was correct, it may have informed policymakers and the broader public about some flaws in those communities. Overall, I'm inclined to mark the factor of producing critiques as neutral: I can see the appeal, but I don't think past critiques were hugely productive.

I'm not sure what to make of their new director. Optimistically, he is an experienced academic who could breathe new impetus into CSER. Besides his successful academic career in the US, he previously "joined a multi-disciplinary team responsible for forecasting long-range threats to critical infrastructure" for the US Department of Homeland Security, which could be the type of practical policy engagement that could make CSER's work valuable in practice. On the other hand, in a previous interview with SoGive, he came across as a bit clueless. However, after some reflection and after an interview with Nathaniel Cooke, I did see some sense to the concept of secrecy risks[3]. Maybe the next interview will be informative.

1.3. Negative aspects of CSER

CSER seems to not be big on prioritization. Per various interviews, with Seán, with an anonymous person close to their group, or with Nathaniel Cooke, they generally seem allergic to the idea. Funnily enough, Seán flagged it as a "conflict of interest" that he was more excited about the AI part of CSER.

This allergy to internal prioritization has the result that they will likely not seek to grow the most valuable parts of CSER differentially. This makes it less desirable to give CSER unrestricted funds.

To illustrate this point, consider Toby Ord's estimates of existential risk in

[3] In short, secrecy might enable existential risks, in the same sense that secrecy enabled the Manhattan project. Doesn't seem crazy to look at, even though it's a bit specific.

The Precipice. He estimates a chance of existential risk in the next 100 years of around 1 in 10k for supervolcanoes[ˆvolcanoes], and around 1 in 10 for unaligned artificial intelligence. In addition, addressing existential risk from supervolcanoes seems tricky and capital intensive, and my sense is that it's not more tractable than risk from artificial intelligence. But CSER has researchers working on both artificial intelligence and on risks from volcanoes, and therefore it seems pretty plausible that the differences in impact between its researchers can be ~1000x or higher.

To elaborate on this point further, one of the researchers for whom CSER was looking for funding was Lara Mani, who specializes in risk from volcanoes. I had Vasco Grillo look into this more deeply; his results are here. He indeed concludes that existential risk from volcanoes is extremely low. To elaborate on this though even more, here are profiles for a large fraction of CSER researchers, elaborated by a researcher from outside the EA community. The picture these profiles give is of an eclectic institute with wildly varying effectiveness profiles.

So a donor can do better by not donating to CSER as a whole, but rather to the parts of CSER the donor identifies as more valuable. As a result, I will be modeling what I consider to be one of such valuable parts, CSER's AI group[4]. I feel that this is straightforward, but also that this was worth pointing out explicitly.

CSER's funding comes from a potpourri of sources, like the Grantham or the Templeton Foundation for climate risks, or the Effective Altruism or Open Philanthropy sphere for risks from artificial intelligence or biological risks. And these sources seem to be aware of the mishmash nature of CSER, and try to avoid funging.

Still, this combination of many funders and a lack of prioritization leads to a lack of intellectual coherency. It will be difficult for there to be a "vision for CSER as a whole", because CSER's resource allocation decisions are made by their funders, not by their leadership[5].

Perhaps that is illustrated by the fact that the pathways to impact I outlined in the previous section aren't from a research agenda for CSER as a whole (there doesn't seem to be one), but rather my best guess. Compare with this pathway to impact for GovAI.

CSER is also pretty expensive. Their cost per researcher is above $100k; $100k for salaries and pension. But this doesn't include their actually pretty large support staff. In contrast, cost per researcher for an organization on the other end of the cheapness spectrum, like Riesgos Catastróficos Globales, is something like $20k.

[4] I will also model some degree of funging between that valuable part and the less valuable part of CSER.

[5] This is maybe a weird way to put it. The converse would be that CSER hasn't gained the trust of funders for unrestricted funding.

After some internal drama and infighting, Seán generally seems tired of the reins of leadership. Understandable, but it also makes CSER as a whole less valuable.

As a final negative, CSER hasn't been particularly responsive to requests, so I actually don't know where their marginal funding would go if left to their choice. I also don't have a great handle on their current director.

1.4. Synthesis: A tricky and potentially exciting opportunity for impact

Combining the positive and negative aspects of CSER, we can characterize as a tricky, and perhaps therefore exciting, opportunity for impact. The hope is that if we set up a donation just right, we might achieve a large amount of impact, in a way which other donors wouldn't because it's too complicated.

Besides making recommendations to its donors, SoGive could also take an activist investor role mediating some possible Open Philanthropy funding[6]. On the CSER side, this would involve giving restricted funding, standing available as an advisor, or making its case about where marginal funding could be most useful. On the Open Philanthropy side, it could make the case to Open Philanthropy that CSER beats their last dollar, and offer to interface with CSER on its behalf[7].

2. Summary of a threshold approach for medium-sized donors

In a previous write-up, we discussed a decision method for choosing whether a donor interested in existential risk should donate to a given opportunity by comparing to a number of thresholds.

The approach was to define some thresholds, and check whether a potential intervention meets them. The thresholds were:

- A minimum willingness to pay threshold based on the notion that the existential risk community should be willing to spend all of its funding to bring existential risk to zero.
- An upper bound on willingness to pay based on the notion that maximum willingness to pay should be some small multiple of global GDP, because higher multiples cannot be paid, i.e., humanity would not be able too coordinate to put in the effort that such multiples would represent.

[6] As a brief aside, activist investor or short-seller funds, like Hindenburg Research, don't typically use only their own funding, but rather get larger funds to invest in them. Here, Open Philanthropy could take the role of the larger investor. Open Philanthropy is attention and capacity constrained, and has "seeing like a state" problems. It wouldn't have the attention and bandwidth to hand-hold CSER much. Perhaps SoGive could seek to explain that to Open Philanthropy or to a larger organization like Founders' Pledge, and get some funding with which to shepherd CSER.

[7] Matthew and Jessica haven't been responsive when I've reached out, though. So I wouldn't be excited about doing this myself. But SoGive might still want to.

- A threshold based on a comparison with some robust funding to the AI safety community, based on the notion that if there was some other reasonably scalable intervention that beat the intervention under consideration, it would be better to fund the better scalable intervention instead.

2.1. Minimum willingness to pay

The most stringent of those thresholds was the minimum threshold, which enthusiastically recommended that an intervention be funded if it provided 1 basis point per $1.4 to $13M. Or, reciprocally, that each million dollars provided 0.075 to 0.67 basis points.

Note that this threshold is a distribution because it has as an input uncertain quantities, like the amount of money in the existential risk community, and the total amount of existential risk. We will consider what difficulties this will bring in section 2.3.

Here are some graphs representing that threshold, as well as some underlying statistics:

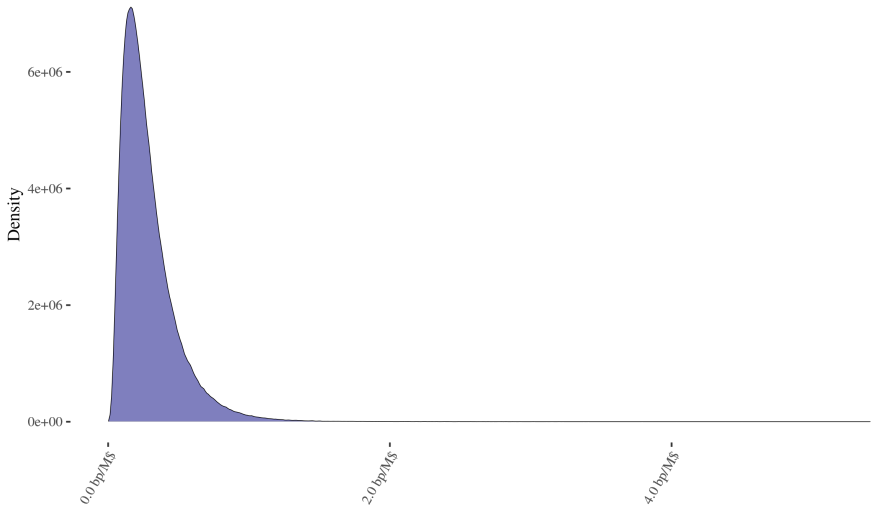


Figure 1: Distribution of the minimum willingness to pay threshold

Statistic	Value
Mean	0.287678
Median	0.237393
Std	0.201160
90% confidence interval	0.075771 to 0.669618

Statistic	Value
80% confidence interval	0.098762 to 0.535675
50% confidence interval	0.151029 to 0.366343

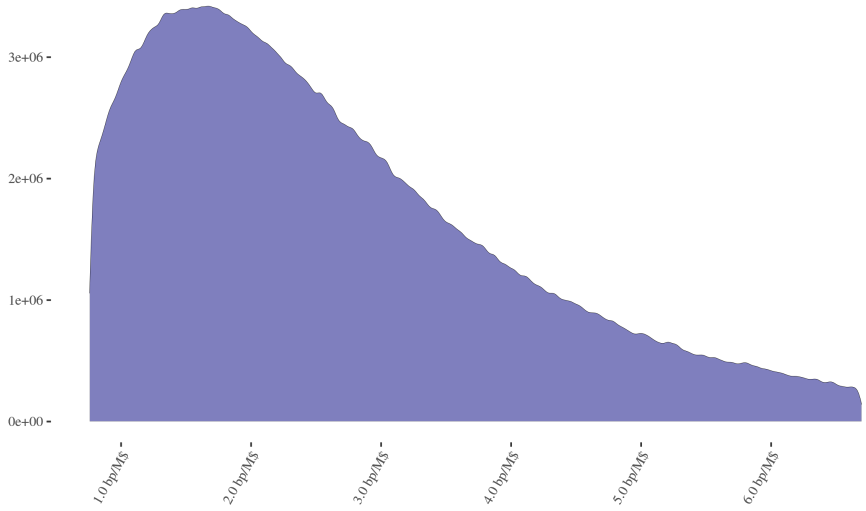


Figure 2: Distribution of the minimum willingness to pay threshold (90% confidence interval)

2.2. A threshold based on not leaving a Pareto improvement on the table

This threshold was based on the estimated value of some robust funding to the AI safety community, and in particular to mechanistic interpretability research, such as MATS. The reasoning was that if there was some other reasonably scalable intervention that beat the intervention under consideration, it would be better to fund the better scalable intervention instead.

Here are some representations for this threshold:

Statistic	Value
Mean	0.135836
Median	0.037410
Std	0.382310
90% confidence interval	0.000802 to 0.554423
80% confidence interval	0.002255 to 0.320412
50% confidence interval	0.009731 to 0.121844

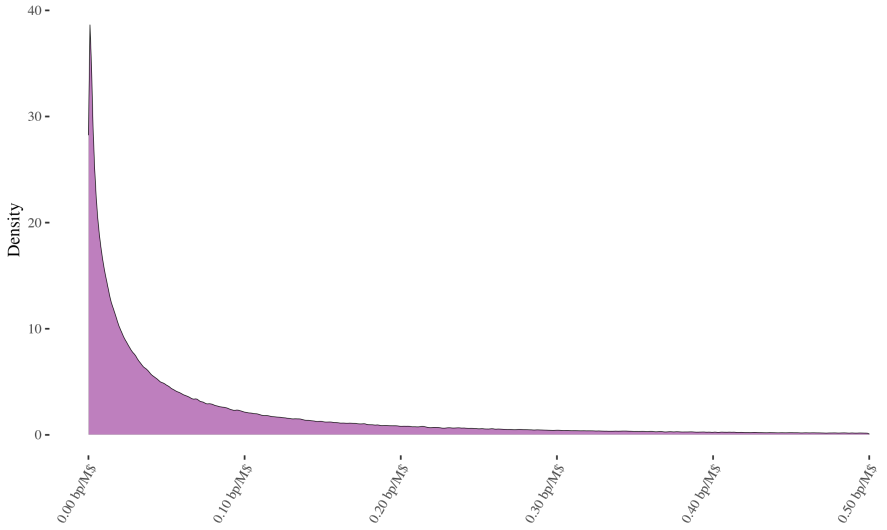


Figure 3: Distribution of the AI community threshold

2.3. Comparison with other distributions and dominance criteria

Potentially, when comparing a possible intervention against these thresholds, we might run into difficulties, because it's not clear when one distribution is better than another one.

Some ways we could compare them could be:

- Compare their means, their expected values
- Compare their 90% confidence intervals
- Compare their medians
- etc.

All of these might be respectable. Personally, I'd be inclined to recommend comparing means.

2.3.1. Dominance criteria

However, we might be tempted to look into dominance criteria. One criterion of dominance is statewise dominance. For instance, in a matrix like the following:

State of the world	Value of A	Value of B
1	10	1
2	200	20

A is better than B in state 1, A is better than B in state 2, and therefore A statewise dominates B, so we should clearly choose A over B.

A similar principle is stochastic dominance, where A stochastically dominates B if, for every amount of value x, $P(A > x) > P(B > x)$. Verbally, this means that you can get more of what you want more of the time. So for example, given these payoffs…

Probability	Value of A
0.3	2
0.7	4

Probability	Value of B
0.4	1
0.6	3

…then A stochastically dominates B, even though it could be the case that A takes a value of 2 and B takes a value of 3.

You can visually see if one option stochastically dominates another one if it looks shifted and scaled to the right.

In our comparison between CSER and the AI community threshold, CSER will stochastically dominate that threshold.

2.3.2. Comparison of means as a comparison of the result of multiple independent draws of similar bets.

Consider again two interventions:

Probability	Value of A
0.5	1
0.5	2

Probability	Value of B
0.9	0.1
0.1	100

In this case, neither stochastically dominates the other one.

However, now consider the distribution of 100 draws of each. They will look as follows:

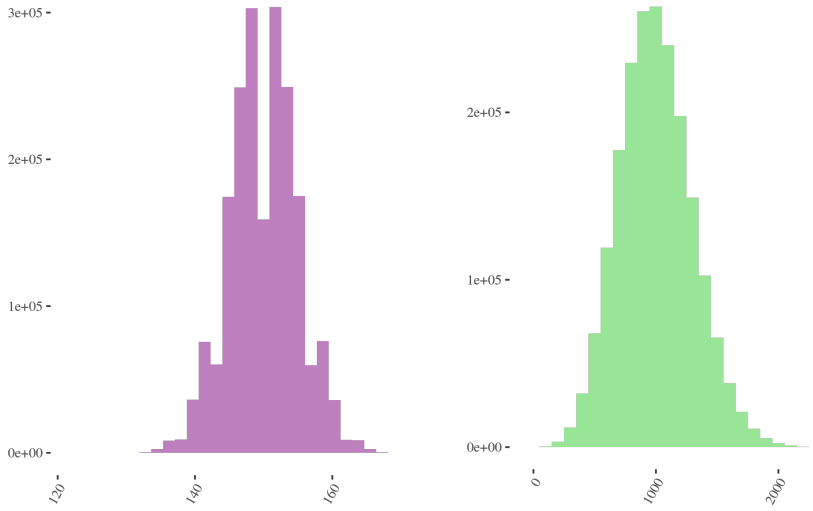


Figure 4: Distribution of the sum of 100 draws from A and 100 draws from B

I mention this because SoGive makes an emphasis in robustness, but some robustness can arise from a community of people trying different things, as opposed to every participant in a community trying to be individually "robust".

When comparing CSER vs the first principles threshold, CSER will have a higher mean due to a larger right tail of impact. And so CSER will beat the threshold in the sense that a bundle of many different bets with its same distribution as CSER will meet the threshold.

3. A model of CSER's impact

3.1. Explanation of the model

The model attempts to estimate the reduction in existential risk. It consists of factors like:

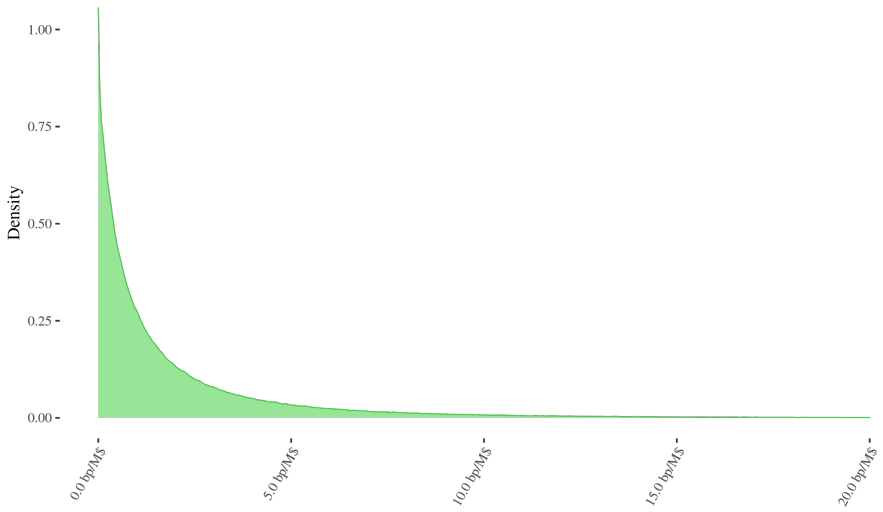
- Total existential risk from transformative risk from AI
- Share of transformative AI research coming from Britain
- Magnitude of AI existential risk coming from Britain
- Magnitude of the reduction in such risk by the AI safety community in Britain
- Share of that reduction assignable to CSER
- Chance of funging with less valuable parts of CSER for future funding
- Influence in other countries as multiplier of influence in Britain
- Value of marginal future funding as a proportion of current value

Ultimately, none of these factors end up being decisive, but rather CSER's value is explained as their combination. Also, even though most of these factors are quite uncertain, the uncertainty over their product ends up being tighter. This happens often when multiplying distributions; for instance, multiplying lognormals will tend to reduce their standard error relative to the mean (for a proof sketch, see the last section here).

You can see the model in more detail here. For instance, here is the small model estimating the share of AI risk coming from the UK. I estimated each of the factors subjectively, looking at a reasonable lower and upper bound based on my understanding of the world, and then fitting a suitable distribution depending on the shape of the uncertainty: usually a longnormal distribution, but also a beta distribution for the case where variables, like probabilities, were bounded between 0 and 1. To fit a 90% confidence interval to a beta distribution, I used this tool.

Initially, I was planning to have two versions of the model, one written in C, and another written in a more accessible language or tool, like Carlo or squiggle. However, when it came to it, it turns out that Carlo wasn't able to "put a distribution inside another distribution", which is necessary for the funging model, which was composed of a) the value of the rest of CSER vs its AI part, and b) the fraction of a donation which is funged. And Squiggle was similar enough to the current model that there is not much accessiblity gained. Instead, for now I've recorded instructions to run the model here[8]

3.2. Result of the model

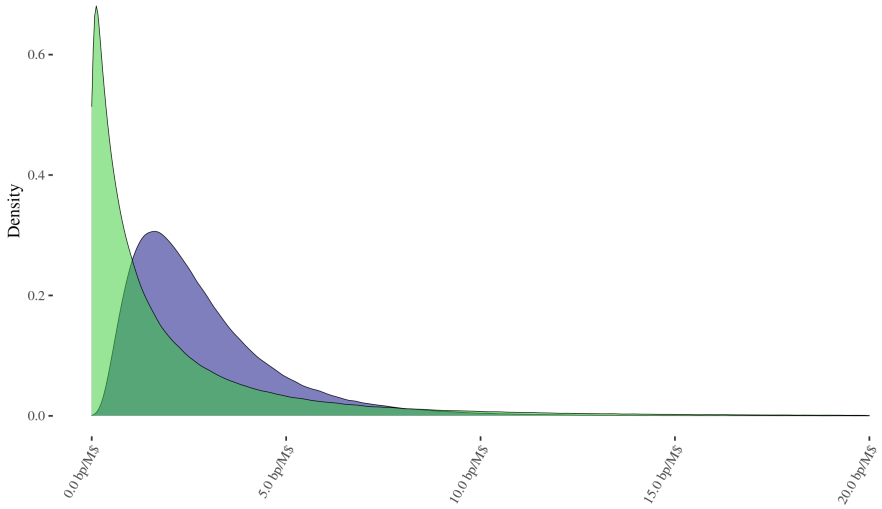


[8]The video is unlisted, but can be accessed with its link.

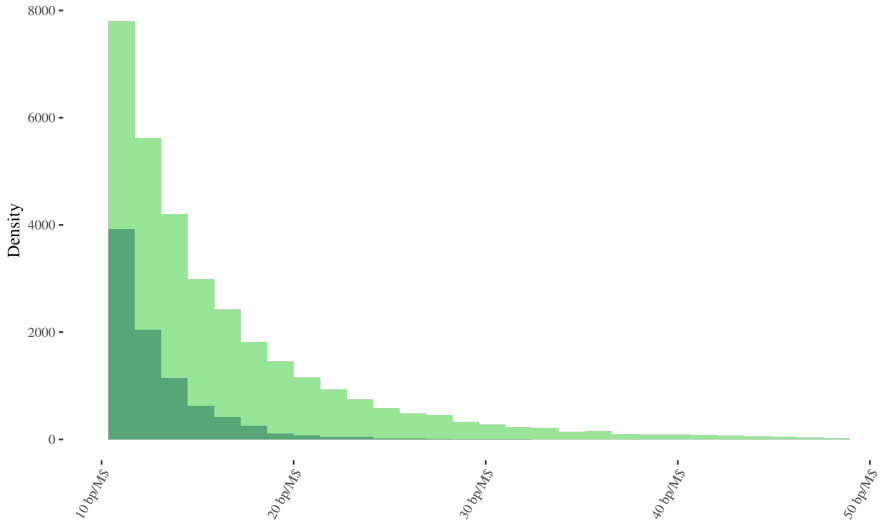
Statistic	Value
Mean	3.174159
Median	1.682010
Std	4.462334
90% confidence interval	0.145405 to 11.196894
80% confidence interval	0.264881 to 7.650863
50% confidence interval	0.665697 to 3.872561

3.2. Results of the model compared to the strict first principles threshold

Statistic	Value
Mean	3.174159
Median	1.682010
Std	4.462334
90% confidence interval	0.145405 to 11.196894
80% confidence interval	0.264881 to 7.650863
50% confidence interval	0.665697 to 3.872561



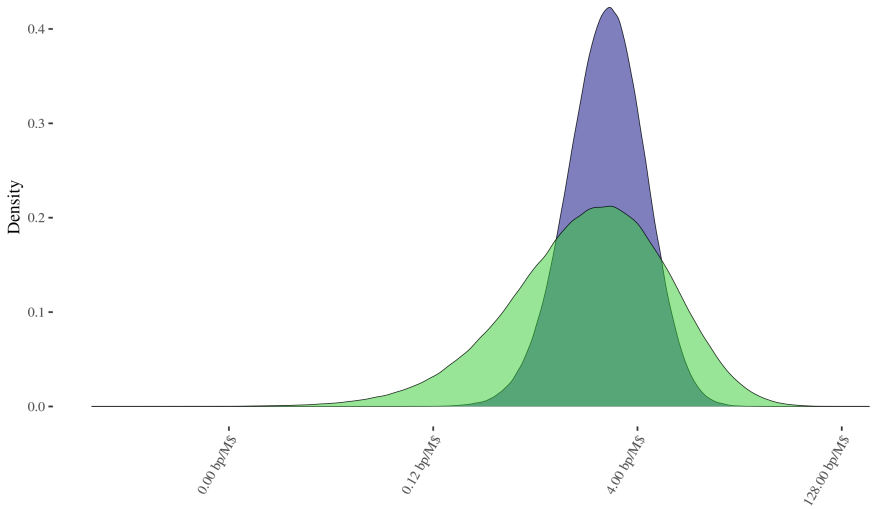
To see the tail, we can zoom in:



or we can plot the x axis on a log scale:

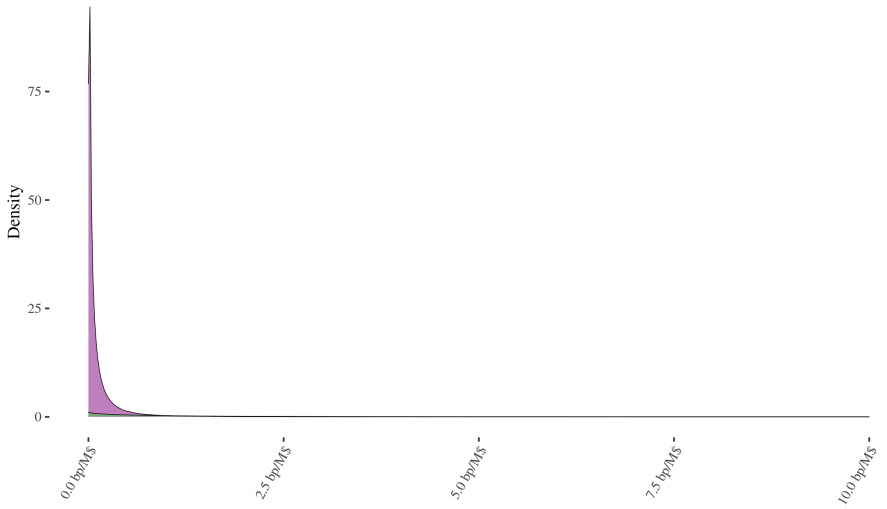
So the threshold initially seems to be higher. But CSER has a higher mean and a longer tail. Given some moderate amount of risk neutrality, or a community making bets like CSER, the threshold would be met.

The argument is a bit clearer if we consider the value of the AI part of CSER without considering funging. If so, the tail is a bit more visible:

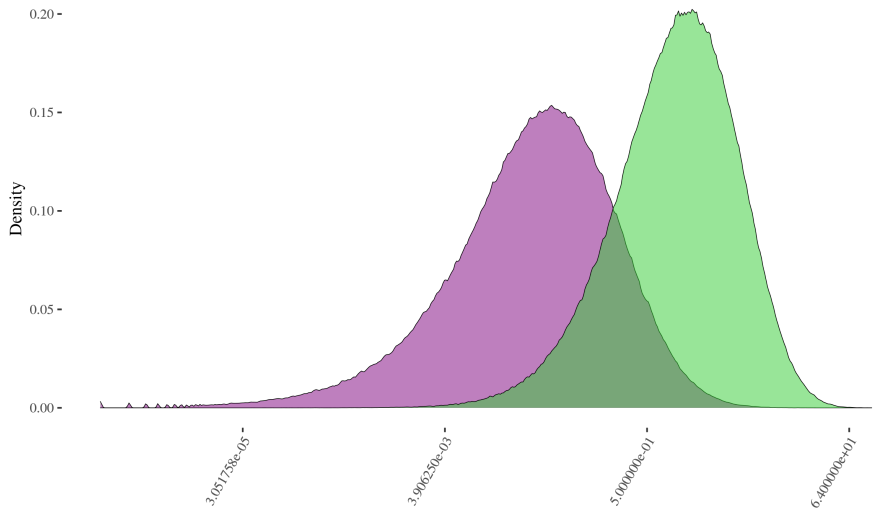


3.2. Results of the model compared to the less strict AI safety community threshold

Statistic	Value of CSER
Mean	3.174159
Median	1.682010
Std	4.462334
90% confidence interval	0.145405 to 11.196894
80% confidence interval	0.264881 to 7.650863
50% confidence interval	0.665697 to 3.872561



or, on a log scale:



So in this case, CSER does stochastically dominate what I modelled as a robust donation to technical AI safety.

3.3. Some commentary about what this means and where CSER is at.

We start looking at the overall chance of AI risk, and we set it at 2%% to 20%%, as a wide interval, some combination of my own subjective estimate and an intuitive aggregate of others' risk estimates.

Then we ask, of that share, what fraction belongs to the UK? The case for very little, i.e., something like 1/20th is: There are many cutting edge labs: OpenAI, DeepMind, Inflection, Anthropic, StabilityAI, Mistral, Facebook AI, Baidu/Tencent/Beijing Academy, Grok. And more are popping up, so it might even be the case that in the future, the best lab is an organization that hasn't been founded yet. On the other end, the case for a lot, 1/4th is that the truly cutting edge labs are OpenAI and DeepMind. And looking at recent job openings for DeepMind, about half are in the UK, so we could divide the influence half and half between OpenAI and DeepMind, and of DeepMind's half, about half goes to the UK.

Then we look at the state of AI risk in the UK, and we notice that actually, the AI risk community seems to be doing a pretty good job, with the AI summit, secondments deep in the government, and an AI institute that is spending real money on safety. So then we assign an impressive, but fairly uncertain 10%% to 80%% safety reduction of the UK's AI risk to the UK's AI safety community. This is much, much better than other peer countries.

We still have to ask, though, what fraction of that reduction in impact should be

assigned to CSER. The case for comparatively little is that you have a bunch of organizations working on the topic, and in particular, GovAI is working directly with DeepMind. The case for comparatively more is that FHI recently suffered several scandals and CSER has been "holding the line", and in general could just be more competent than others. Overall I 7%% to 50%% of the impact of the UK's AI safety community to CSER in particular.

For the last two paragraphs, the interview with Seán was pretty informative. One specific result of that interview was that I also added a multiplier for international impact, mostly for their work with China, but also to a lesser extent for the Vatican, which could end up having some outsized impact. That multiplier is a fraction of CSER's impact in the UK, and I'm estimating it as 5%% to 50%%.

Having established that CSER is doing some valuable work, we look to the value of an additional donation. On the one hand, hiring a new person wouldn't be as valuable as current work. On the other hand, some of the time CSER may have to le t a current person go. For instance, Maathis Maas left them to go to Legal Priorities, a worse organization where he is having less impact. This reduces the impact of a marginal donation a bit.

In addition, because CSER has other areas which aren't as valuable, and it has some unrestricted funding, a marginal donation could substitute for some of that unrestricted funding. This could end up being pretty bad depending on the quantity, and ends up depressing the impact of a marginal donation. Another possible source of funging would be Open Philanthropy, I've relegated discussion of it to an appendix.

Putting these factors together, CSER does beat one of the thresholds outright, the one which models some robust technical AI safety intervention, like SERI MATS. In words, CSER has just had enough policy impact that it blows it out of the water.

But then the comparison against the more ambitious "first principles" threshold is a bit trickier. My model of CSER says that CSER has a higher mean than the threshold. But it overlaps with it. And, because we have multiplied many uncertain factors together, in a big chunk of possible worlds, CSER's impact, considered as an individual bet, doesn't meet that threshold. Still, a series of independent bets like CSER definitely would. The bottom line is that CSER is pretty good, but the domain in which it operates is uncertain, and the model reflects that. My recommendation would be consider the threshold to be met, unless in your heart of hearts you know yourself to be very risk averse.

4. Conclusion

So, CSER dominates the threshold of some robust funding to the AI technical safety community, and it beats the more ambitious threshold based on first principles, depending on how we look at it.

From here, the natural course of action seems to be as follows: seek to reduce funging within CSER, and then make a restricted donation. Another option might be instead to lobby Open Philanthropy to fund CSER. But this depends on your level of risk averseness, if you were very risk averse, you might want to choose some other intervention.

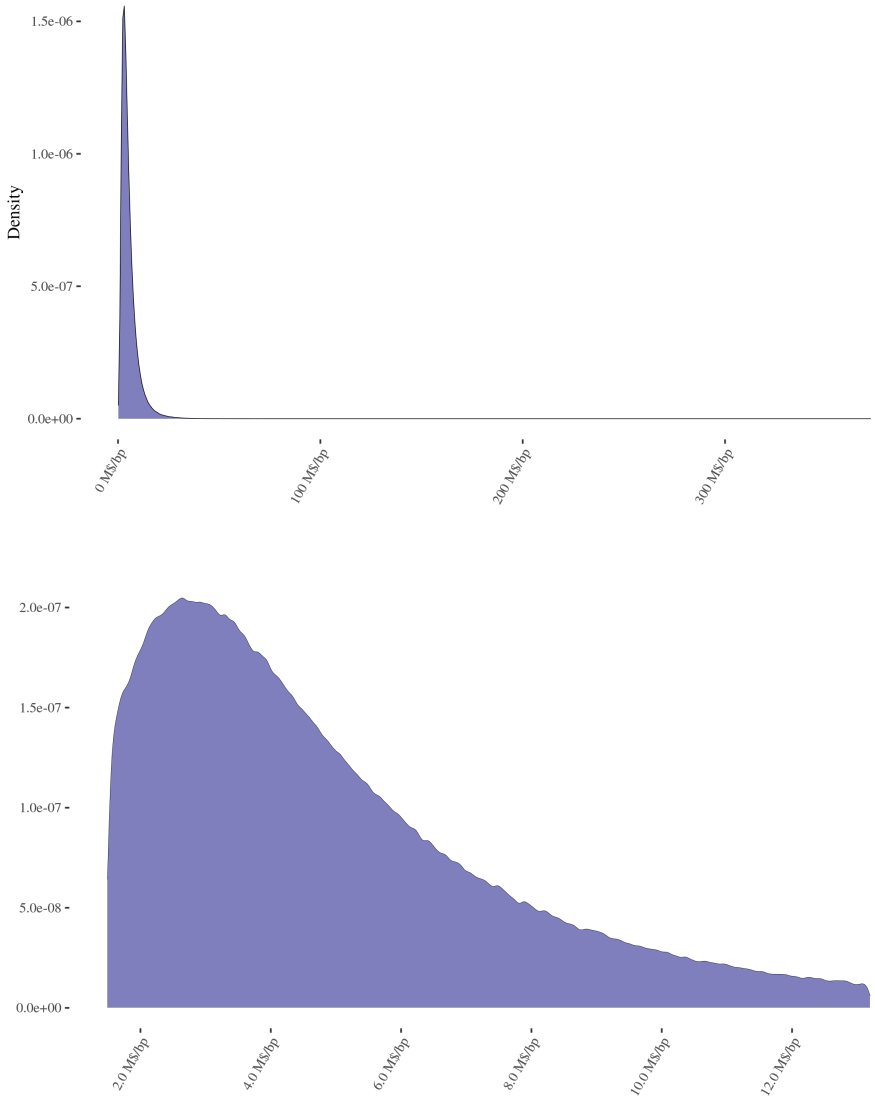
Appendices

§A. Millions of dollars per basis point vs basis points per million dollars.

Previous drafts talked about the minimum willingness to pay threshold in millions per basis point. That has some meaning. But it also divides something (dollars) by a probability in a way which is, in some sense, confusing. Dividing by dollars is a measure of cheapness, but dividing by a probability seems… inelegant.

Anyways, the threshold document talked about M$/basis point, so for continuity, here is that threshold visualized in terms of M$/bp. In those terms, it recommended that an intervention be funded if it provided at least 1 basis point per $1.4 to $13M. Or, reciprocally, it recommended that each million dollars provided at least 0.075 to 0.67 basis points.

Statistic	Value
Mean	5.4 M$
Median	4.2 M$
Std	12 B$
90% confidence interval	1.5 to 13 M$
80% confidence interval	1.9 to 10 M$
50% confidence interval	2.7 to 6.6 M$



§B. Adversarialness

A factor which could change the value of a donation to CSER would be whether it funges against Open Philanthropy, i.e., whether Open Philanthropy would fund a donation to it, and what OP would do with its money otherwise.

Here, I have conflicted feelings. On the one side, I tend to think that OP has a bunch of unique bottlenecks, and that a small donor should be able to do much better, which leads to me wanting to be somewhat "adversarial" towards Open

Philanthropy, i.e., to not take opportunities that Open Philanthropy would otherwise fund.

On the other hand, my impression of SoGive is that it doesn't particularly want to be adversarial towards Open Philanthropy. And also, that SoGive is just starting to evaluate and donate to speculative/longtermist opportunities. That leads me to wanting to postpone the adversarialness until SoGive is more established and more confident that it can in fact find opportunities that beat OP.

For now, I have not included a factor considering funging with OP. But to some extent this is an "executive decision", i.e., something about what you want rather than something about what is the case, so if you want me to include a funging factor against OP let me know and I'll do that.

§C. Robustness and deep uncertainty

The above draft has generally assumed a certain worldview, a way of looking at things, in which existential risk from AI is at least somewhat likely, there exist some policy proposals that reduce such risk, and CSER is capable of identifying them and advocating for them. But, in the best of worlds, existential risk ends up being a nothingburger, and we all live happily ever after.

This leads us back to SoGive's desire for robustness. Maybe one way to achieve some of that robustness would be to allocate some fraction of the funds, say, 10%, to red-teaming exercises. Happily enough, my sense is that CSER would have the capacity to carry out those exercises, given that it has people skeptical of AI risk, or who hold that the community is wrongly going about reducing them.

§D. Extinction vs existential risk

When doing the above modelling, there is a choice to be made between modelling extinction risk or modelling existential risk. Extinction risk is much more more conceptually crisp (literally all humans die). But when I think about this domain, I think my intuitive probabilities refer to the broader and more fuzzily defined concept of existential risk. Here, existential risk is a cluster concept, or which extinction is the prime example. But there are other scenarios which would also fall under "existential risk", e.g., where humanity "loses control", where there is some dark age, or where many but not literally all people die, where humanity is crippled in some important way, or where it ends up in some dystopian scenario.

So I've chosen to refer to existential risk when doing my modelling.