Cost-effectiveness evaluation of artificial intelligence-based existential risk reduction work at the Centre for the Study of Existential Risk

Nuño Sempere

Edited by Spencer R. Ericson on November 13, 2024

Abstract

I present positive and negative aspects of CSER, and then present a synthesis of CSER as an opportunity for impact. I model its impact and apply a threshold model. I find that CSER's AI work beats a strict threshold for a risk-neutral donor, averting about 3.17 basis points of existential risk per million dollars donated (mean estimate). I discuss modelling choices and caveats, and conclude with two recommendations: seek to reduce funging within CSER, and then make a restricted donation to their AI work.

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 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. Overview of CSER

1.1. Introduction and positive aspects of CSER

The Centre for the Study of Existential Risk, CSER, is a 23-person 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
  - To do this as part of a whole existential risk community, because many voices have more credibility than only one voice
  - To organize workshops; recently also send participants to secondments
  - To be more left-wing than right-wing; it's possible this could pay off within the Labour government[1]
  - To be affiliated with Cambridge, which is useful for credibility for policy-makers
  - To be present in the UK, which is a bet on the UK's relevance in AI; the AI summit validates that bet so far

Per an interview with Seán Ó hÉigeartaigh[2](CSER's past Director), 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 policies increases robustness, which has paid off after the successive FHI scandals. Affiliation with Cambridge leads to increased cost, but also to closeness

[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 the Labour government.

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

to policy impact. My sense is also that CSER might be in a better position, politically, since Labour has won the UK election.

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 concerns (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 saw democratization as 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: they have potential but haven't paid off enough yet.

1.3. Negative aspects of CSER

Per various interviews with Seán, with an anonymous person close to their group, and with Nathaniel Cooke, CSER seems to prefer treating their programs equally, instead of determining which projects are highest-impact. Seán flagged it as a "conflict of interest" that he was more excited about the AI part of CSER.

This attitude towards 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 *The Precipice*. He estimates a chance of existential risk in the next 100 years of around 1 in 10k for supervolcanoes, 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 Grilo look into this more deeply. He indeed concludes that existential risk from volcanoes is extremely low. Looking at the profiles of other CSER researchers, CSER appears to be an eclectic institute with widely 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 the most valuable parts, CSER's AI group[3]. 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. These sources seem to be aware of the varied nature of CSER, and try to avoid funging. Thus, it will be difficult for there to be a "vision for CSER as a whole", because CSER's resource allocation decisions are made from outside the organization, by their funders, to the extent that funding is restricted.

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 publicly), 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, with $100k just 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.

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

Combining the positive and negative aspects of CSER, we can characterize this 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[4]. From SoGive to CSER, this would involve giving restricted funding, standing available as an advisor, or making its case about where marginal funding could be most useful. SoGive could make the case to Open Philanthropy that CSER beats their last dollar, and offer to interface with CSER on its behalf.

[3] I will also model some degree of funging between the especially valuable AI work at CSER and the parts of CSER that I expect to have less impact on existential risk.

[4] 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.

We will now move on to quantitatively comparing CSER's AI work to thresholds for cost-effectiveness.

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. We will review this write-up now.

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.
- 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.
- A maximum 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.

Threshold	Mean value in M$/bp	Median value in M$/bp	Std in M$/bp	Mean value in bp/M$	Median value in bp/M$	Std in bp/M$
Minimum willingness to pay threshold	5.41	4.21	4.46	0.185	0.237	0.224
Robust AI technical safety work threshold	9.38e4	26.7	7.04e7	1.07e-5	0.0374	1.42e-8
Maximum willingness to pay threshold	6.59e3	3.16e3	1.21e4	1.52e-4	3.17e-4	8.25e-5

2.1. Minimum willingness to pay

The rest of Section 2 is technical notes on how to compare interventions to thresholds. Skip to Section 3 for the analysis of CSER's AI work.

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 (90% CI). 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:

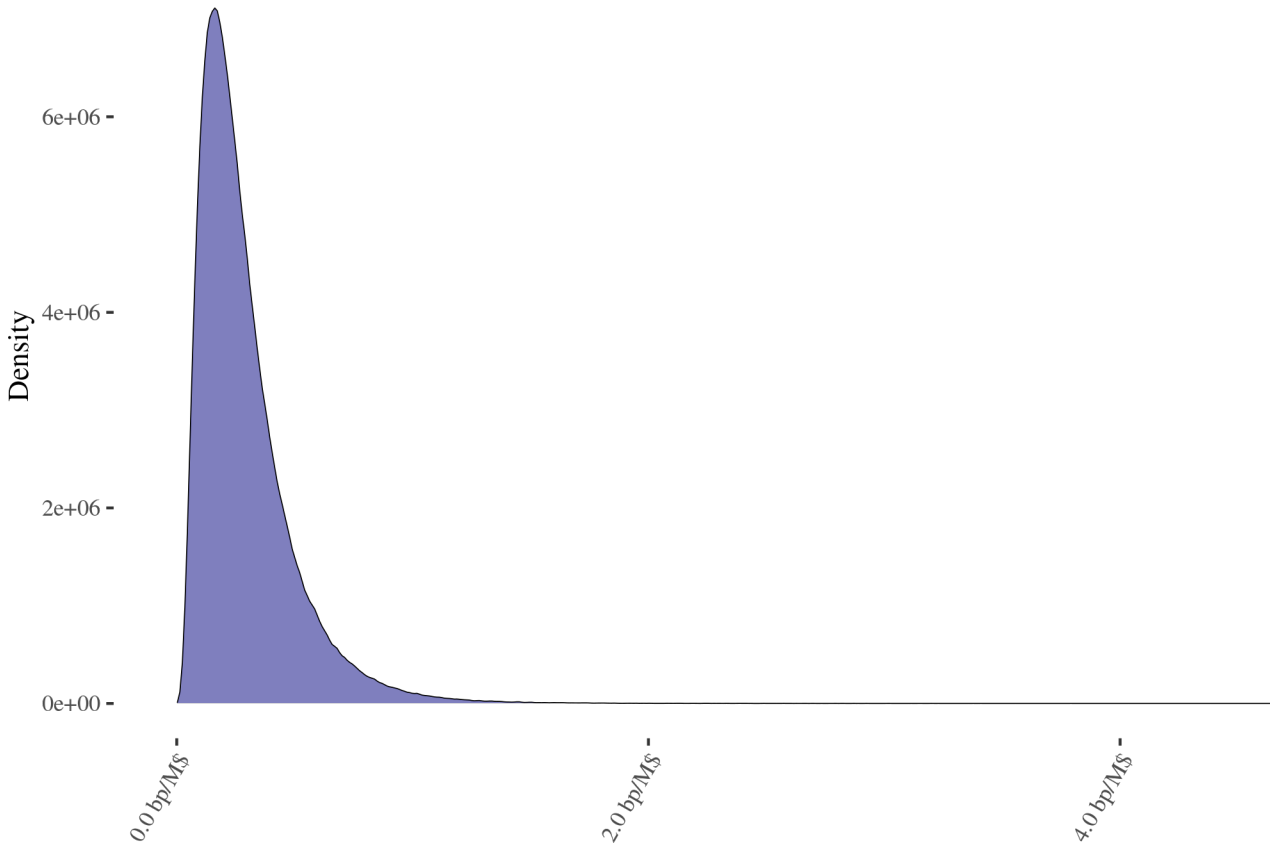


Figure 1: Distribution of the minimum willingness to pay threshold

Statistic	Value (bp/M$)
Mean	0.287678
Median	0.237393
Std	0.201160

Statistic	Value (bp/M$)
90% confidence interval	0.075771 to 0.669618

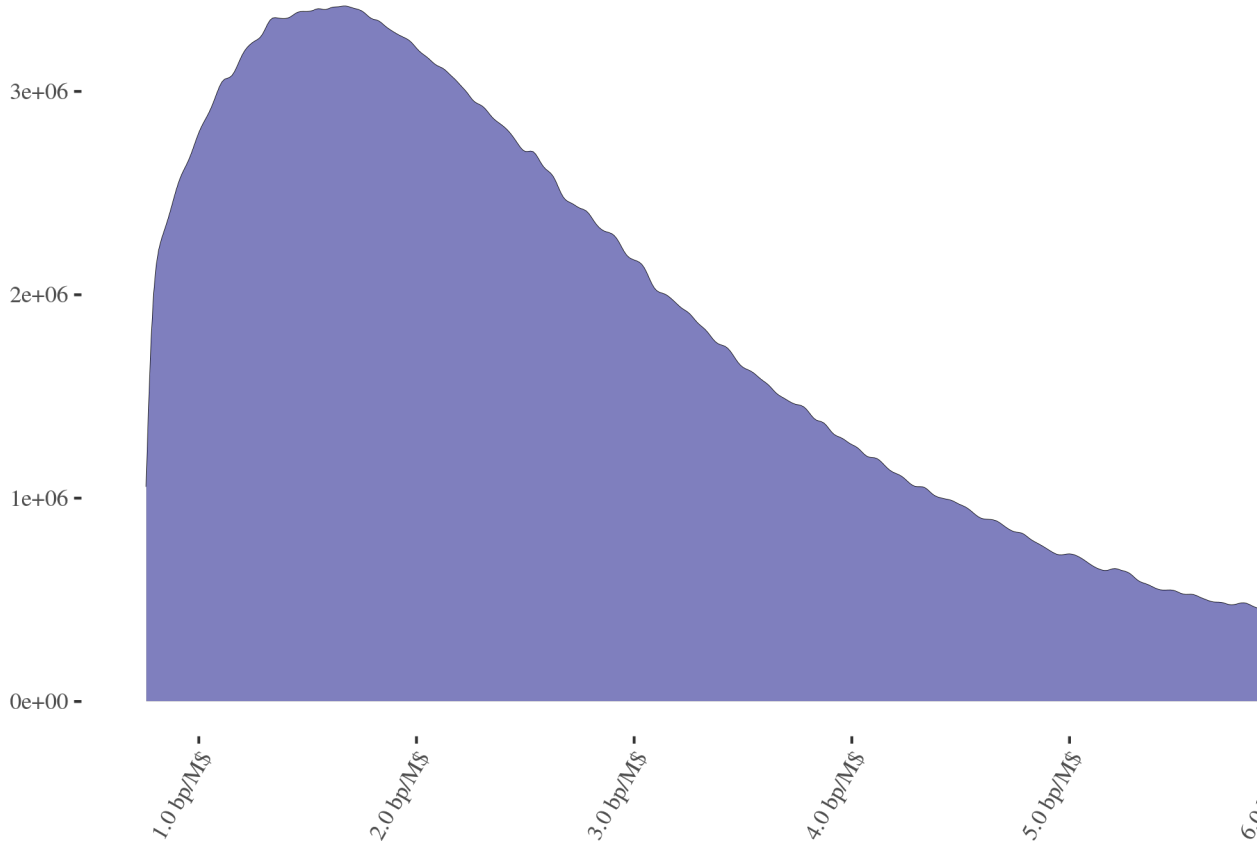


Figure 2: Distribution of the minimum willingness to pay threshold (90% confidence interval)

2.2. A threshold based on the value of AI safety technical research

This threshold was based on the estimated value of some robust, scalable funding to the AI safety community, for instance some mechanistic interpretability research (which probably has a very small downside) at MATS. We can use a robust, scalable program like MATS as a benchmark. 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 more scalable intervention instead.

Here are some representations for this threshold:

Statistic	Value (bp/M$)
Mean	0.135836
Median	0.037410
Std	0.382310
90% confidence interval	0.000802 to 0.554423

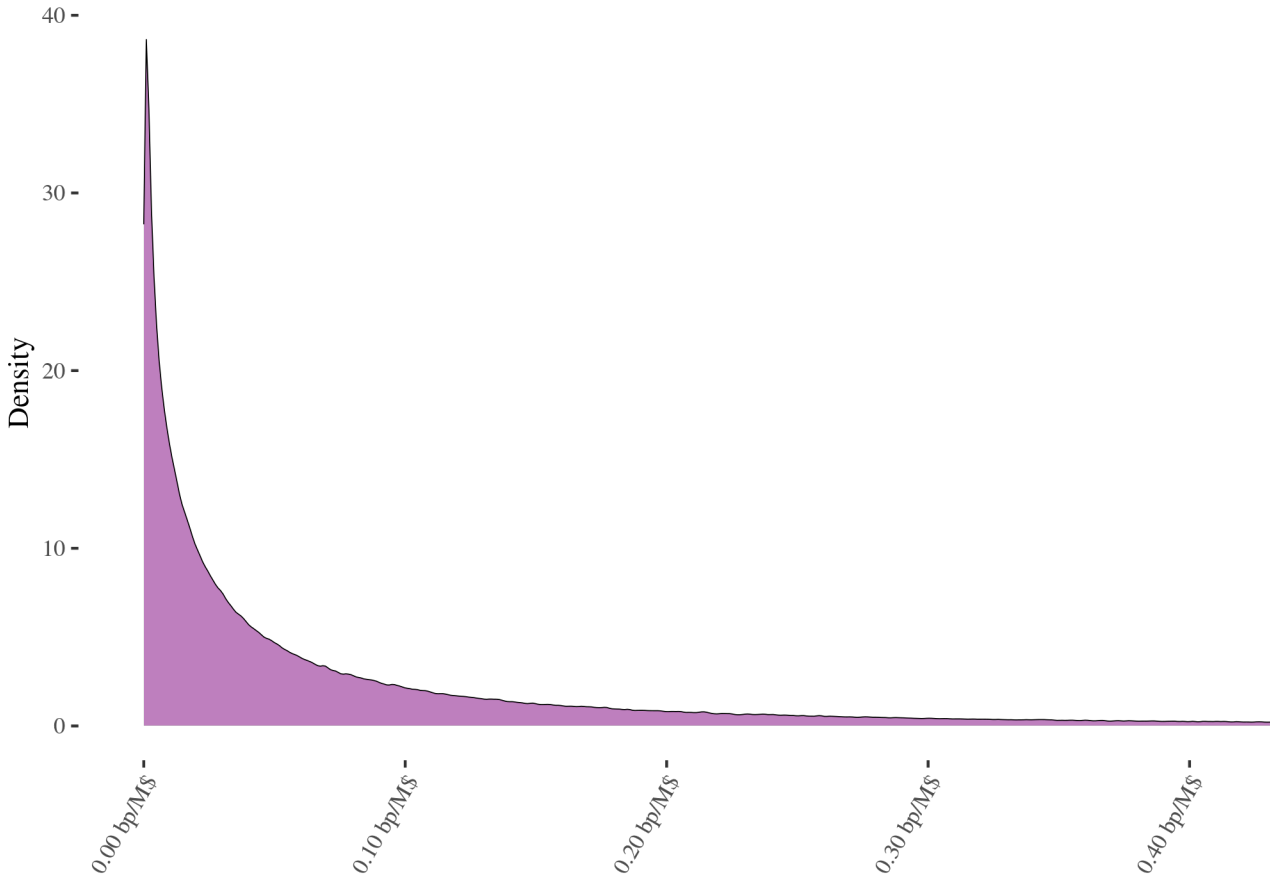


Figure 3: Distribution of the AI community threshold

2.4. A maximum willingness to pay

The third threshold had a value of $400M to $23B per basis point, and it was meant to more quickly discard possible interventions. For instance:

- Consider the possibility of avoiding $10e^{-7}$ of existential risk from volcanoes[5] for $1B. This would have a cost per basis point of $100B, and the threshold wouldn't be met.

[5]I'm choosing this number as the midpoint between Toby Ord's 1 in 10e4 and Vasco Grillo's 3 in 10e10 estimates of existential risk from volcanoes per century. Considering distributions doesn't change the picture.

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, i.e., 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, we will see that CSER stochastically dominates 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. Figure 4 shows how this will look.

SoGive makes an emphasis on robustness. 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".

We will see that when comparing CSER to 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 the same distribution as CSER will beat the threshold.

3. A model of CSER's AI efforts.

3.1. Components of the model

- 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 preferred parts of CSER for future funding
- Reduction in existential risk in other countries as multiplier of reduction of existential risk in Britain
- Value of marginal future funding as a proportion of current value

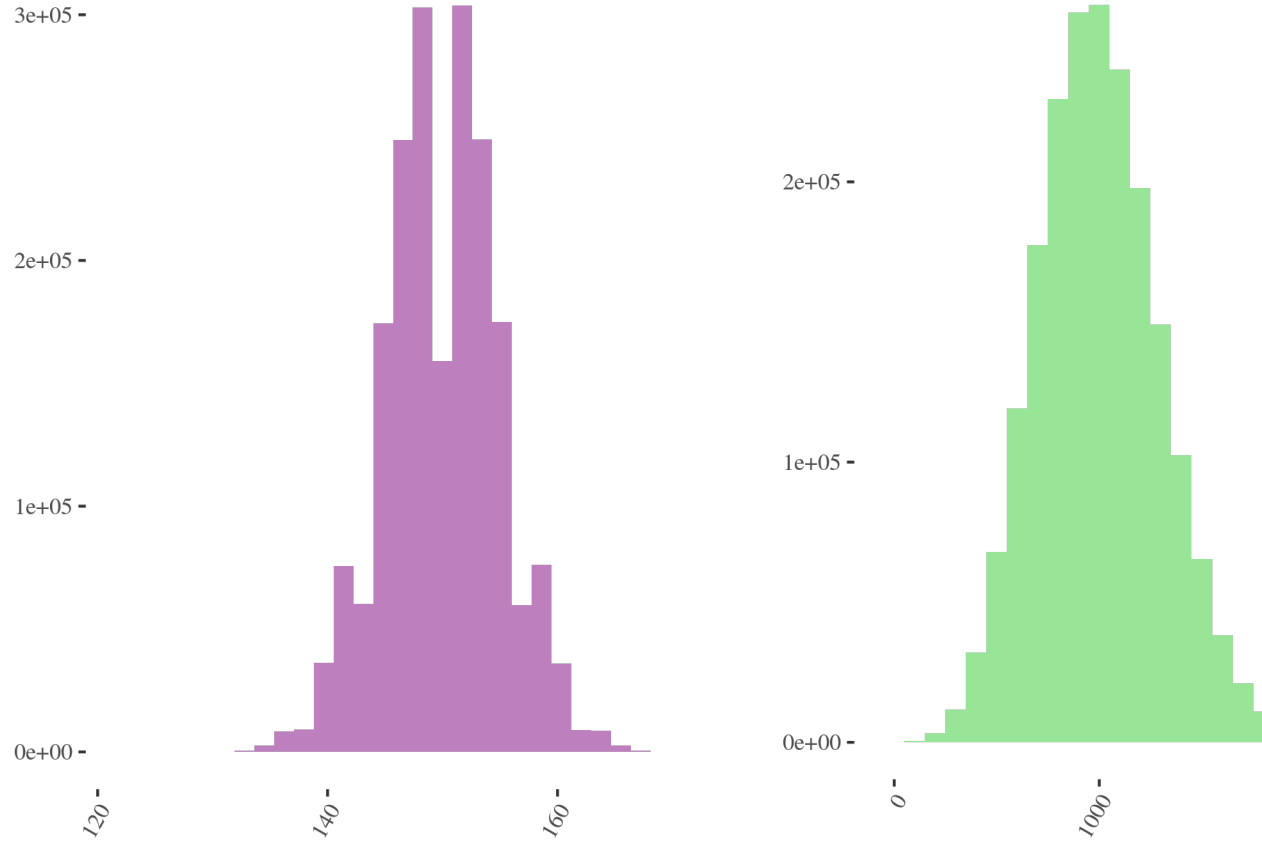


Figure 4: Distribution of the sum of 100 draws from A and 100 draws from B

For more methodological notes, see Appendix E.

You can see the model in detail here. Generally, 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 lognormal 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 on my website.

3.2. A narrative overview of the model of CSER's AI work

We start looking at the overall existential risk from AI. 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. (Editor's note: Some reasonable readers might put the lower bound several orders of magnitude lower. This iteration of the model is therefore most useful to a donor whose internal models of existential risk are close to this. We encourage readers with significantly different assumptions, on this quantity and others, to fork the model and see how their assumptions change the cost-effectiveness of CSER's AI work.)

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 hand, the case for a lot (1/4th) is: 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.

Here is the small model estimating the share of AI risk coming from the UK. (Editor's note: SoGive has uncertainties here and would especially appreciate the readers' efforts to make this quantity more clearly defined.)

Consider the state of AI risk in the UK. We notice that 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. We assign an impressive, but fairly uncertain 10% to 80% reduction of the UK's AI risk to the UK's AI safety community. This is much, much better than other peer countries. (Editor's note: Some of the work in the UK improving AI safety is not philanthropically funded work, such as DeepMind's internal efforts. SoGive is unsure whether this factor should lower the credit here, in the context of how the thresholds for cost-effectiveness were drawn.)

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 assign 7% to 50% of the impact of the UK's AI safety community to CSER in particular. (Editor's note: SoGive would appreciate readers' insights on whether other organizations in the UK would have pressed on with AI safety work productively in the absence of CSER. We are currently sympathetic to the view that CSER's work has enabled the work of other actors, but it's possible that the lower bound should be lower

here.)

For the value of the UK's AI safety work and CSER's role in it, my interview with Seán was 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%.

We then 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 let a current person go, and there is an impact cost here that funding could prevent. For instance, Maathis Maas left them to go to Legal Priorities. Potentially, with more funding, such fellows could have stayed, or transitioned to more senior positions. Overall, this reduces the impact of a marginal donation a bit.

We then consider the chance of funging. CSER has some unrestricted funding, and a donation to CSER could substitute for some of that unrestricted funding, such that that donation ends up funding a non-AI part of CSER. This 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 Appendix B.

Ultimately, none of these factors end up being decisive, but rather CSER's value is explained as their combination. 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).

Initially, I was planning to have two versions of the model, one written in C[6], and another written in a more accessible language or tool, such as a spreadsheet. The funging model is not as accurate in the spreadsheet model because of the nested distributions. However, you can find the spreadsheet model here for accessibility of reproducing the results. This spreadsheet can be fed into Carlo here to visualize the results nicely.

[6]These are brief instructions to run the original model in C, suitable to a technical audience who is interested in the additional accuracy this model brings.

3.3. Result of the model

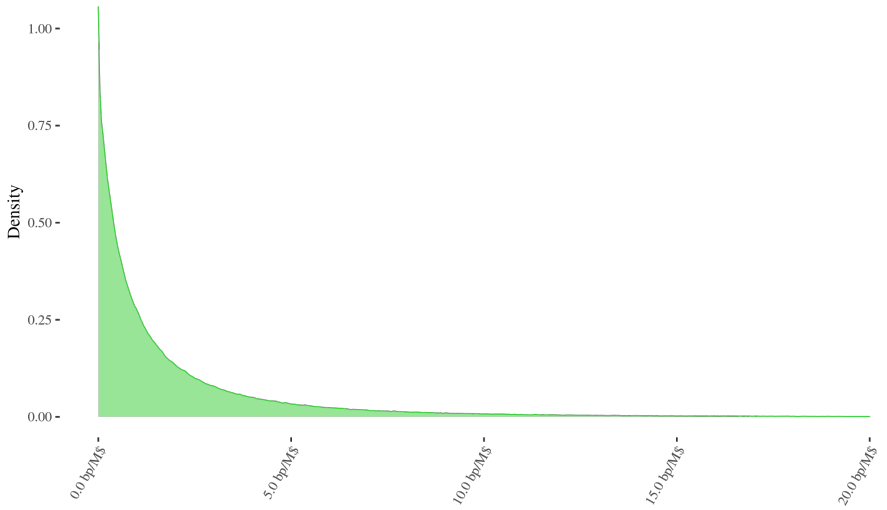


Figure 5: Distribution of the reduction in existential risk by CSER's AI program, in basis points per million USD in 2024

Statistic	Value (basis points per million USD)
Mean	3.174159
Median	1.682010
Std	4.462334
90% confidence interval	0.145405 to 11.196894

(See the outputs file on Github here.)

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

Statistic	Value (bp/M$ from CSER AI)	Value (bp/M$ from minimum willingness to pay threshold)
Mean	3.174159	0.287678
Median	1.682010	0.237393
Std	4.462334	0.201160
90% confidence interval	0.145405 to 11.196894	0.075771 to 0.669618

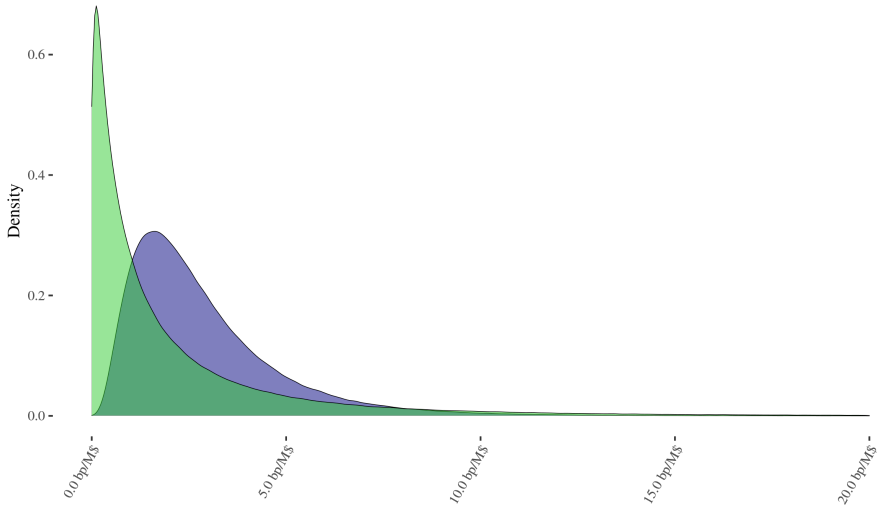


Figure 6: Distribution of the reduction in existential risk by CSER's AI program in light green, overlaid with the distribution of the minimum willingness to pay threshold in blue

To see the tail, we can zoom in:

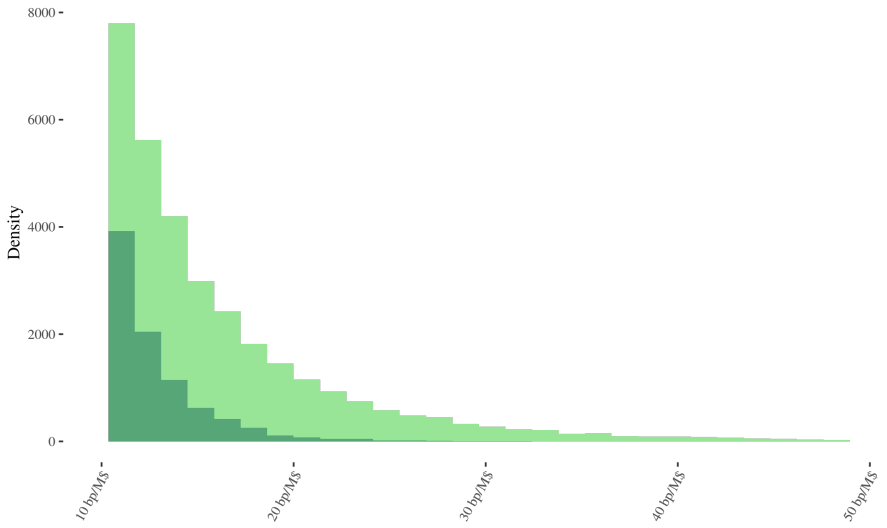


Figure 7: Right tails of the distribution of the reduction in existential risk by CSER's AI program in light green, overlaid with the distribution of the minimum willingness to pay threshold in blue (appearing here as dark green

from transparency)

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

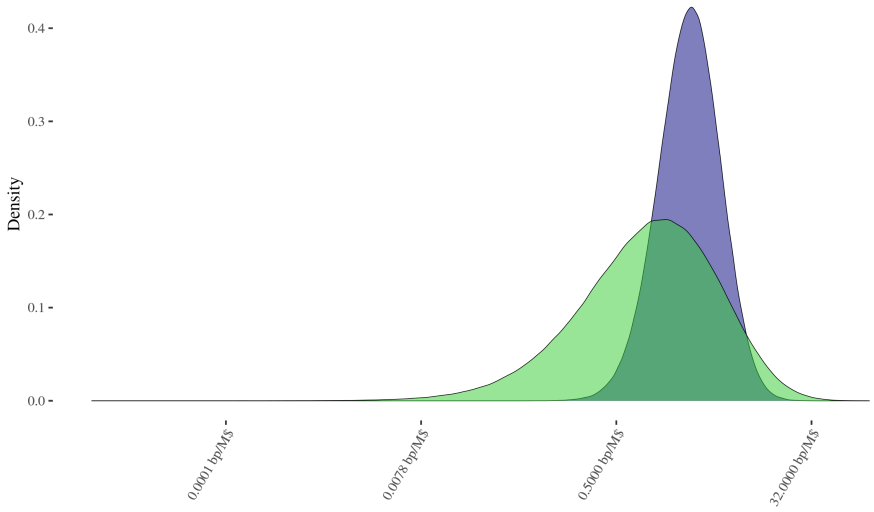


Figure 8: Distribution of the reduction in existential risk by CSER's AI program in light green, overlaid with the distribution of the minimum willingness to pay threshold in blue, with basis points per million USD shown on the x-axis in a log scale

The AI program at CSER has a higher mean and a longer right tail than our minimum willingness to pay threshold. Given some moderate amount of risk neutrality, or a community making bets on organizations like on the AI part of CSER, the threshold would be exceeded.

Note that the default model includes some default chance of some funging between different parts of CSER. If we don't include that funging, the value of donating to CSER AI has a more visible tail and looks even better:

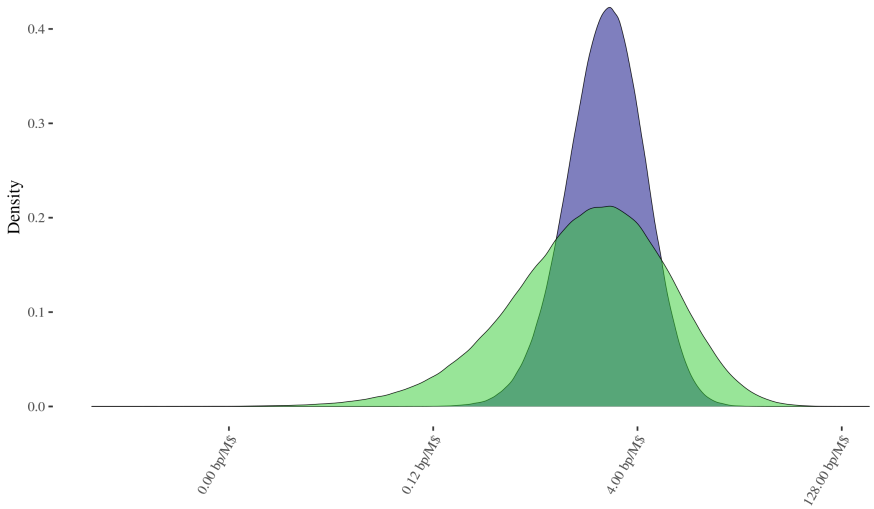


Figure 9: Distribution of the reduction in existential risk by CSER's AI program *assuming no internal funging* in light green, overlaid with the distribution of the minimum willingness to pay threshold in blue

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

Statistic	Value (bp/M'$ to CSER AI)	Value (bp/M$ from AI safety community threshold)
Mean	3.174159	0.135836
Median	1.682010	0.037410
Std	4.462334	0.382310
90% confidence interval	0.145405 to 11.196894	0.000802 to 0.554423
80% confidence interval	0.264881 to 7.650863	0.002255 to 0.320412
50% confidence interval	0.665697 to 3.872561	0.009731 to 0.121844

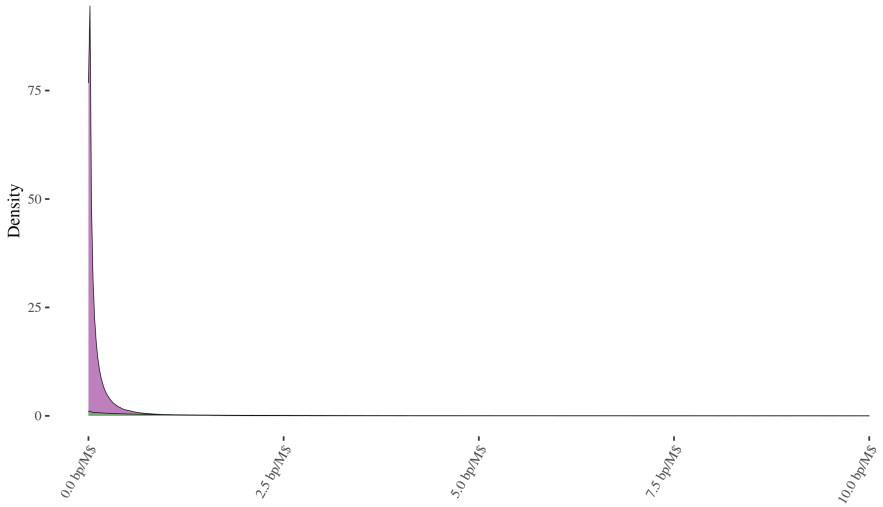


Figure 10: Distribution of the reduction in existential risk by CSER's AI program in light green, overlaid with the distribution of the AI safety community benchmark threshold in purple

or, on a log scale:

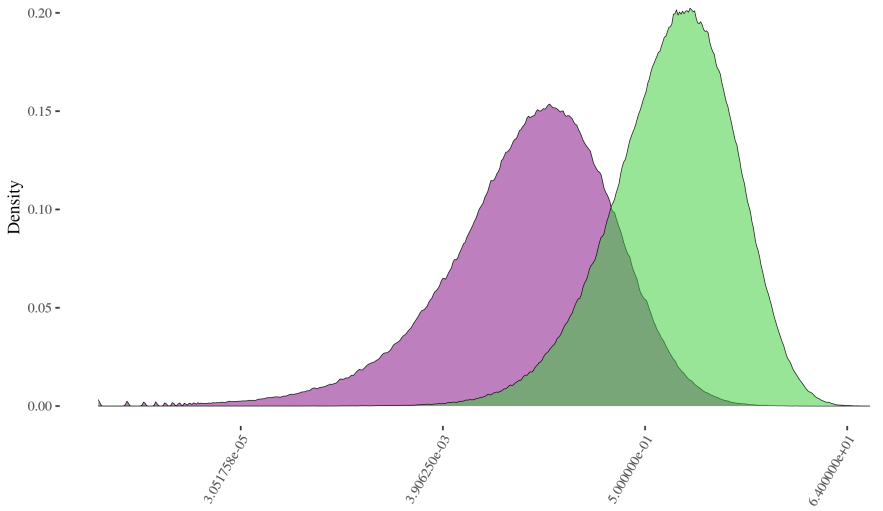


Figure 11: Distribution of the reduction in existential risk by CSER's AI program in light green, overlaid with the distribution of the AI safety community benchmark threshold in purple, with basis points per million USD plotted on

the x-axis on a log scale

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

3.6. What this means

Putting these factors together, CSER outright beats a robust technical AI safety intervention, such as what might be produced by MATS. CSER has had enough policy impact that it blows it out of the water.

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 in many possible worlds, CSER's impact wouldn't meet that threshold as an individual bet, because we have multiplied many uncertain factors together. This is relevant to a risk-averse donor.

But for the risk-neutral evaluator like SoGive, a series of independent bets like CSER definitely would exceed the threshold. 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 to SoGive would be to consider the threshold to be met, since SoGive is risk-neutral.

4. Conclusion

In a previous document, we compared strategies for estimating the value of speculative interventions and deciding whether to fund them. This time around, we gave a brief overview of one particular strategy, thresholds. Then, we compared CSER, and in particular the AI safety policy part of it, against those thresholds. We presented a few methods of doing that comparison: an extremely strong statewise dominance criterion, a strong statewise dominance criterion, and a normal or weak comparison of means criterion.

CSER's AI program ends up beating one of the thresholds outright, the threshold comparing it to some low-downside AI technical safety intervention, as might be exemplified by a graduate student doing technical interpretability work as part of the MATS program. In my modelling, CSER has had enough policy impact that it blows it out of the water.

The comparison against the more ambitious "first principles" threshold (i.e., "minimum willingness to pay") is a bit trickier. My model of CSER says that CSER has a higher mean than the threshold. For expected value maximization, beating the mean is enough.

A series of independent bets like CSER's AI program would, considered together, stochastically beat the ambitious threshold.

Statistic	Value (bp/M, $CSERAI$)|$Value(bp/M$, minimum willingness to pay threshold)	Value (bp/M$, AI safety community threshold)	
Mean	3.174159	0.287678	0.135836
Median	1.682010	0.237393	0.037410
Std	4.462334	0.201160	0.382310
90% confidence interval	0.145405 to 11.196894	0.075771 to 0.669618	0.000802 to 0.554423

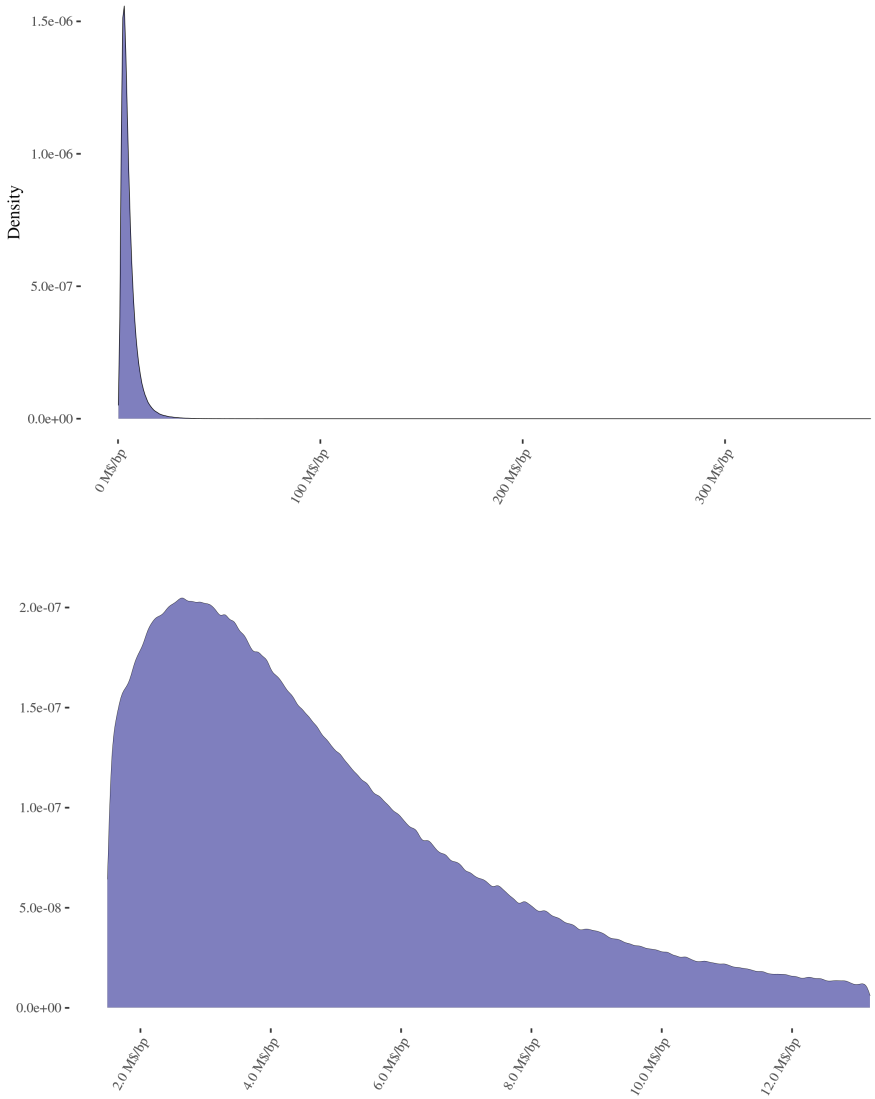
From here, the natural course of action seems to be as follows: seek to reduce funging within CSER, and then make a restricted donation to CSER's AI work. Another option might be instead to lobby Open Philanthropy to fund CSER's AI program.

Appendices

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

The threshold document talked about M$/basis point, although it's an inelegant unit – so for continuity, here is the "minimum willingness to pay" threshold visualized in terms of M$/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	4.5 M$
90% confidence interval	1.4 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".

§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, of 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.

§E. Evaluation of other parts of CSER

A general approach for estimating the value or cost-effectiveness of CSER's work in some area is:

1. Estimate total existential risk for an area (e.g., AI risk, bio risk, nuclear risk, risk of an asteroid impact, etc.)
2. Estimate how much the field of people dedicated to preventing one particular risk have reduced the magnitude of that risk
3. Estimate the share of impact of the field that belongs to CSER
4. Multiply the above and divide by CSER expenses

A key part of step #2 is to outline some accomplishments by CSER, and compare them to the rest of the field.

Throughout this post, I've applied that model to the AI policy part of CSER, because I thought that was the part of CSER that had the best chance of beating the threshold we originally settled on. I've also applied that approach, though more superficially, to risk from volcanoes, and argued that it didn't meet a threshold for funding.

Here are brief sketches of how to apply that estimation method to other parts of CSER:

- Bio: Estimate the chance of a catastrophic pandemic, the reduction in that probability from advocacy movements, and the share that belongs to CSER.
  - On the bearish case: this isn't a hugely neglected field, because there is a lot of interest after COVID-19. Also, CSER's bio team is relatively small and hard to scale, and doesn't seem to have major wins.
  - On the bullish case: There is still massive interest after COVID-19, perhaps leading to large opportunities, and perhaps CSER could be well positioned to advocate for changes in Britain using that interest.
  - Note that in conversation, Matthew expressed the view that they were bottlenecked by *mentorship* on bio, so currently marginal funding to bio seems low-impact.
- Nuclear risk: CSER has very few levers for affecting major sources of nuclear risk, which are going to be actions and policies of state-level actors like Russia or Iran.
  - As a result, they let go of their one nuclear policy person, Paul Ingram.
- Applying this decomposition to uniquely synthetic, "holistic," or "systemic" risks seems more of a stretch, but one could also attempt to estimate what magnitude of risk comes from such risks, and what share of that does CSER prevent.

Other parts of CSER are likely to be substantially less valuable than its AI policy part (to a risk-neutral donor who intends to reduce existential risk). A donation would be more valuable if it avoided funging with them. One variant that I'm worried about is the superficial rebranding of some other thing (risk communication, ethical problems of mathematics, conundrums in foresight) into "AI".