Towards Comprehensive Existential Risk Assessment: A Bayesian Network Model And Proposal For Assessment

Siebe Rozendal[1]

26 - 8 -2019

10 pages

Epistemic status: Here, I outline a proposal to improve existential risk assessment. I am more confident that better x-risk assessment is needed than that my approach is the best approach. It is only my current best guess and very much subject to change.

Note October 8th, 2019: My current priority is to find a promising PhD position where I can do research related to this project. However, the topics discussed here might be too overarching to be doable in the current academic environment. Thus, I am unsure how much of this project I will take forward, and this is not a highly polished document.

Abstract

To effectively reduce existential risk, we need a clear understanding of what the biggest and most urgent risks are. To answer this, we must first answer: **how should we assess existential risk?** Current estimates are either non-existent, unsubstantiated, incomplete, lack information, fail to adequately communicate uncertainty, or suffer from other methodological weaknesses such as selection bias. This paper presents a Bayesian network model to assess existential risk, focusing on two primary outcomes: extinction and non-recovery from societal collapse. The model has several theoretical contributions. First, it discounts the existential risk of a specific hazard by the probability that a catastrophic trajectory change occurs due to other potential hazards and by the probability that *societal invulnerability* is achieved. Second, it introduces *aggravation factors* to quantify how a hazard might change the status quo trajectory even without the hazard leading directly to collapse or extinction. After describing the most relevant parameters and their relations, a mixed methods approach is proposed to assess the values of the parameters and their associated uncertainties. The mixed methods approach includes interviews and workshops to uncover hidden cruxes and use distributed expertise, as well as surveys to reach a wide and representative group of experts. Finally, the paper presents tentative implications for existential risk reduction and identifies important topics for further research.

[1] This is the result of a Summer Visitorship at the Centre for the Study of Existential Risk at the University of Cambridge. I am grateful for feedback from Seth Baum, Simon Beard, Haydn Belfield, Zoe Cremer, Max Daniel, Luke Kemp, Matthijs Maas, Seán Ó hÉigeartaigh, and Luisa Rodriguez.

Introduction

Many actors who try to reduce existential risk currently act on a very incomplete strategic picture. Few comprehensive surveys have been done, with the exception being Sandberg & Bostrom (2008), which was an early 'quick and rough' survey.[2] . This survey lacks methodological rigor, does not provide justification for the claims made, fails to communicate the uncertainty in each estimate, and most certainly does not represent a diversity of views. Although many actors surely have more complex and nuanced views on existential risk strategy and prioritization, this is not yet represented in a comprehensive risk estimate. Many disagreements seem to remain within informal discussions. On the other hand, more sophisticated methods to estimate risks exist and have been used (Cf. Baum, De Neufville, Barrett (2018), Diaconeasa et al. (2018) for estimates on the probability of nuclear war). However, many focus only on estimating a (component of a) single risk or do not attempt to capture the diversity of views.

One tradeoff in current estimates is that they are either 1) about the total existential risk but highly subjective and unfounded, or 2) better founded, but limited to a specific scenario and hard to compare with other risks. The goal of this project is to strive towards a comprehensive existential risk estimate that is decently justified and can be improved over time.

The model presented here builds on Baum et al. (2018), in which the authors identify different possible trajectories for human civilization. Specifically, the model is a Bayesian network that requires estimates of the probabilities of the different possible paths to existential catastrophe for each major hazard (cf. Tonn & Stiefel (2013) who recommend this method and identify alternatives). This would identify the relative likelihoods of different risk paths and help to identify at which points intervention seems most promising. Additionally, surveying experts on these probabilities would help to surface important disagreements (cruxes) and locate where our uncertainty is high.

The model

The current model focuses on only two types of existential risk: extinction and non-recovery from global societal collapse. The plan is to expand to other types and even suffering risks once a good method has been established for these risks. When only these x-risks are considered, the formula looks as follows:

$$p(X) = p(Ext_{direct}) + p(Col) * (1 - p(Rec|Col)) \; + p(Col) * p(Rec|Col) * p(X|Rec)$$

$p(X)$ *Total existential risk* when only human extinction and non-recovery from civilizational collapse are included.

[2] This survey does not seem to inform the field's current priorities, nor does the field see the survey as representative or accurate. Nevertheless, it's one of the only comprehensive estimates that have been made which rely on more than the judgment of one or two people.

$p(Ext_{direct})$ Probability of *direct extinction*. Currently defined as '100% of humanity dies within one year, and no descendants to humanity remain alive'.

$p(Col)$ Probability of *global societal collapse*. Currently defined as 'loss of the following attributes within 10 years or less:

- At least 50% of global economic activity (Gross World Product)
- World population between 50% up to (but not including) 100%
- Political systems (e.g. none of G20 are able to raise taxes or uphold rule of law)
- Industry

$p(Rec|Col)$ Probability of *societal recovery* to some level, either to the recovery of industry, or digital computing. [3]

$p(X|Rec)$ *Existential risk to a post-recovery society*.

Taking this basic model, we can apply it to each major hazard. To illustrate, it is applied to the risk posed by an exchange of nuclear weapons (*NX* in the formula below).

$$p(X|NX) \approx p(Ext_{direct}|NX) + p(Col|NX) * (1 - p(Rec|Col)) + p(Col|NX) * p(Rec|Col) * p(X|Rec)$$

Note, however, that this is not yet complete; we should also take into account how humanity's prospects get changed by a nuclear exchange that does not lead to collapse. We add to the formula the following:

$$+ p(\neg Col|NX) * p(X|(\neg Col \ \& \ NX)$$

The following figure shows this visually:

[3] The point where a society is recovered is difficult and ambiguous to establish. Recovery of industry is the most useful point, as anthropogenic existential risk greatly increases after the recovery of industry. However, the definition of 'recovery of industry' needs further disambiguation.

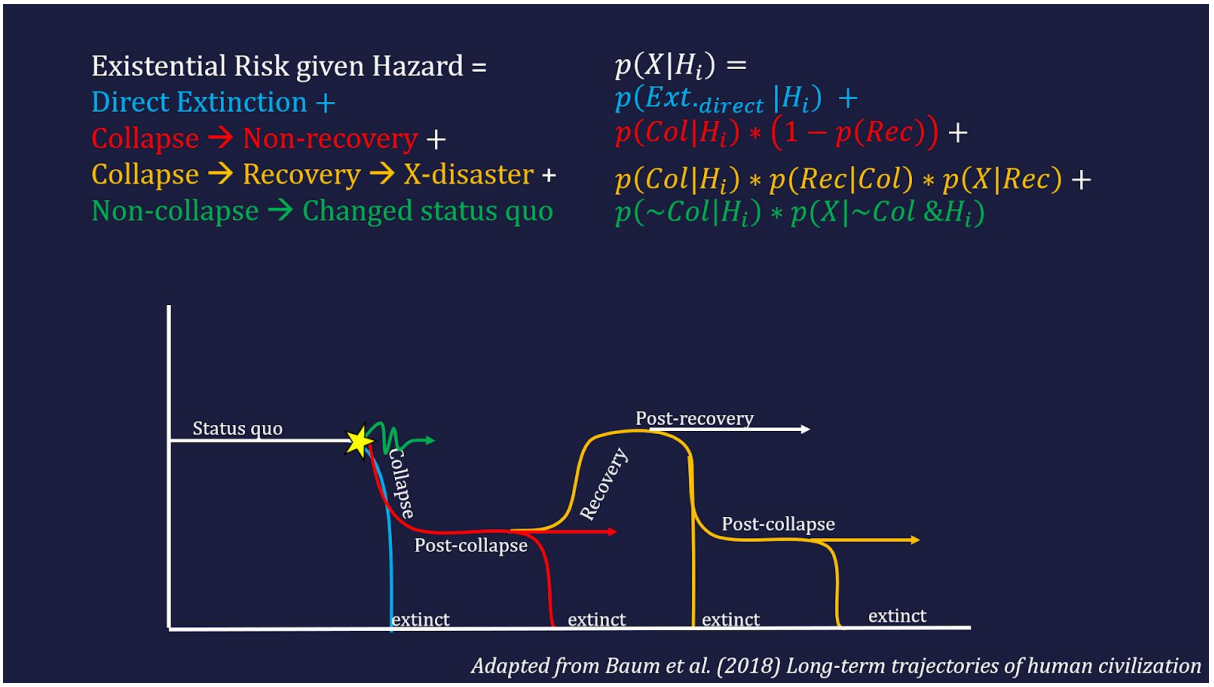

Figure 1. Different paths to extinction, non-recovery from collapse, and recovery

We can also do this for the following hazards. For a first complete estimate, the focus will be on the risk from a pandemic/biorisk $p(Bio)$, AI catastrophe $p(AIC)$, and climate catastrophe $p(CC)$. However, the existential risk of each hazard has so far been estimated with the assumption that *nothing else happens*. Instead, the risk of each hazard should be discounted by the probability that some drastic change occurs.[4] This includes the probabilities of direct extinction, of civilizational collapse, and of achieving 'societal invulnerability' (i.e. the probability that we reach a civilizational competence that reduces x-risk to infinitesimal, for example by achieving beneficent superintelligence):

$$1 - p(Ext_{direct}|\neg h_i) - p(Col|\neg h_i) - p(TM)$$

This can be a substantial factor and can be viewed as the *quantification of urgency*: it does not make sense to focus on far-off risks when the near-term risks are high (or when all problems are expected to be solved soon). This factor appears to point towards a major crux rather than a minor formality; some believe high-level machine intelligence could come soon, which removes the need for dealing with climate change, which has the most severe consequences only in the 22nd century. On the other hand, some will believe civilization collapse is due soon; focusing on far-off technological risk from AI will make little difference

[4] To illustrate, imagine there are two hazards A and B that each have a risk of 50% to lead to extinction in the next 100 years conditional on the other hazard not occuring. The total extinction risk is (obviously) not 100%; instead, it is 1-p(A&B)=1-p(A)*p(B)=0.75. This (posterior) probability is equally distributed over p(A)and p(B), resulting in a posterior probability of .375 that either hazard leads to extinction within 100 years.

This all becomes more complicated when we relax the assumption that the risk is uniformly distributed over time. If instead, A's risk is completely located in the first 50 years and B's risk in the second 50, their posterior probabilities become p(A)=.5 and p(B)=.25 with no change in the total extinction risk. Thus, ideally we ask for a probability distribution over time rather than the probability of X having occurred by 2120.

on this view. However, these considerations become highlighted most strongly when we use probability distributions over time.

Estimating the parameter values

We can take a variety of approaches to estimating each of these variables. They can be split up further into their components. For example, nuclear exchanges can be divided into low total yield (< 25 Megaton, approximately the winter-safe limit of 50 warheads set in Baum (2015) and high total yield ($\geq 25\ Mton$). The probabilities can be triangulated by using different models, and they can be estimated by surveying a diverse set of experts.

I want to present one specific approach to estimating $p(X|Rec)$ and $p(X|(\neg Col\ \&\ NX)$ by introducing the concept of *aggravation factors*.

$$p(X|Rec) = p(X_c) * f_1$$

$$p(X|(\neg Col\ \&\ NX) = p(X_c) * f_2$$

In which $f_i\ \varepsilon\ (0, \infty)$ and $p(X_c) = p(Ext_{direct}) + p(Col) * (1 - p(Rec|Col))$, i.e. the probability of our current global society ending in existential catastrophe. A factor of $f_i = 1$ means the situation is just as good or bad as our current prospects, while a factor of $f_i = 0.5$ means it has gotten much better, and a factor of $f_i = 2$ suggests it has become twice as bad. A more accurate way would also include the proportion of x-risk that is actually affected by changes in the system.

This approach is attractive because it separates the estimate of 'how much better or worse the situation becomes' from the estimate of 'the existential risk after the event'. This allows the estimates to be comparable with other people's estimates without sharing the whole model. If we did not do this, someone who is optimistic about the base rate of x-risk would estimate $p(X|Rec)$ to be low, even though they might believe humanity's prospects would be much worse than our current prospects. In turn, this allows for modularization of the model: we can ask experts for input on only parts of the model.

See here for a Guesstimate model of this formula.

Theoretical issues with the model

A number of issues remain with this model. First, the independence condition does not hold between all hazards. For instance, climate change increases competition over resources, and increases migration. In turn, this affects the likelihood of great power wars that could escalate to a nuclear conflict. However, it could also prove beneficial as the incremental and urgent nature of climate change could lead to effective global governance. Another issue of non-independence is that the occurrence of some events provides evidence for the future probability of other events. For example, someone might infer that a limited nuclear exchange provides evidence of pre-existing international tensions; these tensions could worsen the prospects for cooperation on safe development of advanced technology, which increases existential risk. Although the person has rationally concluded that the observation

of a limited nuclear exchange should increase our credence about existential catastrophe, this does not imply that reducing the risk of limited nuclear exchanges becomes more important. For the moment, non-independence issues will be ignored and independence will be assumed.

Second, we can elicit two different kinds of probabilities: temporal or Markovian. Temporal probabilities are the probabilities over time as used so far. Markovian probabilities are the probabilities that we *eventually* end up in a particular state. They are better suited than temporal probabilities to assess $p(Rec|Col)$ and $p(X|Rec)$ because there is a large uncertainty about the relevant timescales. For instance, humanity could be stuck in a post-collapse state for a 100 years or 10,000 years, but what we care about most is whether humanity recovers *eventually*.

Third, the probabilities depend on the definitions of the different possible states. While the definition for collapse is somewhat reasonable and non-ambiguous, the definition for recovery is harder and needs improvement. Relatedly, not all existential risks fit the paradigm in which one event leads to a major decrease in societal complexity (Cf. Liu, Lauta, & Maas (2018) & Kuhlemann (2019) for 'boring catastrophes'). To take climate change as an example again, much of its existential risk comes from increasing the vulnerability of the system or increasing the risk of other hazards. I plan to do further work of fitting climate change and other slow catastrophes into this model.

Lastly, the model currently does not extend to other existential risks (such as astronomical expansion without a good average quality of life), nor does it differentiate between the different types of existential risk; some states are worse than others (i.e. hellish scenarios) and should therefore be avoided with an even greater degree of priority. These complexities can be added to a later version of the model.

Assessment methodology

The theoretical part of the model has value by itself: it provides structure to the problem and helps important actors to think through the relevant questions. However, given the breadth of the questions, no individual can be expected to have accurate beliefs on each question. Therefore, the distributed knowledge of experts needs to be used. There are many ways in which this can be done; approaches can differ among at least the following dimensions:

- From purely quantitative (e.g. survey asking for estimates only) to purely qualitative (e.g. interviews, scenarios),
- From disconnected (e.g. one-time survey) to highly interactive (e.g. workshops)
- From static (e.g. one-time survey) to dynamic (e.g. improvements over time)
- From comprehensive (e.g. everyone estimates everything) to modular (e.g. experts focus on their own domains)
- From general questions (e.g. estimate probability of collapse) to specific (e.g. estimate probability of this type of scenario)

In the current plan, most of the assessment will focus on small-group workshops (4-6 people) following the IDEA protocol, which is an alternative to the Delphi method.[5] Each

[5] The IDEA protocol is chosen over the Delphi method because the latter focuses on achieving consensus. This is not necessary for the current project and does not improve the epistemic quality of

workshop will focus on a subset of the complete model to allow for depth. A workshop will contain the following steps (based on Hanea et al., 2017, and Hemming et al., 2018):

1. *Introduce*. The goals of the workshop are described and participants are given test questions which can be used to test the logical coherence of their judgments.
2. *Investigate (I)*. Individuals assess the questions individually and come up with their estimates.
3. *Discuss (D)*. Results are given to the group anonymously, and the group is given the chance to discuss the results face-to-face. The discussion is mildly moderated.
4. *Estimate (E)*. Individuals adjust their estimates based on new insights.
5. *Aggregate (A)*. The results will be represented anonymously. They might be aggregated, but only when this is justifiable (cf. Morgan (2014) for discussion).

A fifth step is added:

5. *Reflect*. Participants are invited to share feedback on the process and the contents. This will allow for the process and the model to be improved over time.

The assessment will not use a survey. Surveys are very efficient at collecting answers, but inadequate when questions cannot be answered quickly and instead require deep thought and analysis. Live elicitation in either a workshop or interview can encourage experts to think through all the relevant factors and uncertainties. For this reason, some semi-structured interviews will also be conducted to elicit probabilities. Especially initially, these interviews will help to improve the questions to optimally use the time of multiple experts in workshops.

A note on numbers

Although this model is quantitative, I do not expect this method to yield robust estimates in the near future. It is difficult to calibrate estimates well. Therefore, I think much of the value comes from providing an underlying theoretical framework to guide the thinking of domain experts. Moreover, asking for quantification makes (the amount of) disagreement explicit. A drawback of using quantification is that some experts will actively resist the quantification of belief in domains of high uncertainty. This is reasonable but problematic, because if their beliefs are excluded there is a selection bias towards those with more precise (though not necessarily more accurate) beliefs, which could give the false impression that there is certainty among experts.

Project challenges

I foresee at least the following challenges:

Insufficient expertise on existential risk

Risk assessment via structured expert elicitation requires a group of respondents that cover all the important viewpoints on a topic with a high level of expertise. It is possible that this does not yet exist. In fact, existential risk might be a field in which true expertise can never

the method. Furthermore, face-to-face discussion is often more efficient than online, and the evidence that this biases the process towards groupthink is scant when consensus is not required.

be achieved. The consequences of extreme circumstances on human behaviour may be too chaotic to predict (cf. Baum (2018) for a study of the human consequences of an asteroid strike). In addition, the long timescales involved increase the difficulty of prediction and prevent timely feedback to validate and improve predictions. If expertise does not (yet) exist, it is questionable whether the results of 'expert elicitation' are useful. Morgan (2014) discourages the use of structured expert elicitation on topics that involve predicting complex human behaviour. However, as Bolger & Rowe (2014) note, policy often cannot wait for scientific certainty. This is evident for existential risks. If some information is better than no information at all, structured expert elicitation is very valuable. Assuming this is the case, the project will use structured expert elicitation and present the results as tentative, incomplete, and with other necessary caveats.

Selection bias

Any expert survey is subject to selection bias. It is hard, if not impossible, to define who is an expert and who is not. Can a representative sample of experts be selected? And how to deal with the fact that some expert groups (e.g. climate scientists) are much larger than another (e.g. existential risk researchers)? Selection bias alone might be reason enough to dismiss any survey in areas with high amounts of disagreement between experts. There are three approaches to deal with selection bias: quantitative solutions, qualitative solutions, and accepting that the results are biased. Quantitative solutions include weighing the expertise of an expert (e.g. per class of questions), and clustering worldviews to increase representativeness. Qualitative solutions focus on paying extra attention to experts with diverging viewpoints and identifying their reasons (as done in the Delphi method), or report not averages but clusters of worldviews. Lastly, one can accept that the results are biased and explicitly mention this in any presentation of the results. However, policymakers and the general public are unlikely to pay heed in the reporting of these estimates.

Calibration (on range estimates)

Although domain experts know a lot about their area of expertise, many experts do not know how to properly translate their beliefs into quantitative estimates. Therefore, experts need to be instructed and trained to be better-calibrated. However, there exist no calibration exercises that guarantee proper calibration on long time-horizons because predictions for the long-term future and for unprecedented events cannot be tested. Furthermore, to capture the amount of uncertainty of estimates, the project will use both point and range estimates. Using range estimates creates an extra challenge for calibration, because many people interpret them differently. Furthermore, the philosophical literature on imprecise credences rejects the use of confidence intervals. The current plan is to ask experts for their best point estimates, as well as provide a range by asking something like "if you would have x amount of knowledge, how could your estimate differ from your best point estimates?"

Different models of existential risk

The current project proposal employs one model of existential risk. Having each expert answer the same questions creates a fragile epistemic situation; if a question is wrong, meaningless, or incomplete, a lot of information is so as well. The project should allow for experts to approach the problems from different angles, while still striving for *comparability*

between models and beliefs. The most feasible way to this seems by doing semi-structured interviews that allow for exploration of different approaches and paradigms.

Rigor/effort tradeoff
The questions in the model are high-level questions that depend on many underlying considerations and paradigms. Quick estimates from experts will likely not be very accurate, nor very informative. However, more rigorous and informative estimates will take time and effort from experts. This creates the challenges to make sure the time and effort are well-spent, and to convince the experts of that.

Scope and feedback
Like any other project, this project runs the risk of having too large a scope; assessing the biggest existential risks to humanity is no small task. Additionally, there is a risk of spending too much time and effort on preparation without testing a simple proof of concept. I plan to do interviews, workshop pilots, and small and partial survey pilots to allow for iterative improvement and course correction.

Conclusion

Existential risk reduction requires strategy, which needs to be informed by rigorous x-risk assessment. The current assessment practices are lacking or non-existent. This proposal is to move forward on the basis of a concrete model in combination with a mix of other methods. The path forward is based on the following rough goal hierarchy:

1. Reduced existential risk
   ↑ ↓
2. Improved strategy
   ↑ ↓
3. Good existential risk assessment
   ↑ ↓
4. Understand reasoning of experts and their cruxes
   ↑ ↓
5. Improve assessment methodology

The project will start by trying to get good estimates and derive strategic implications for x-risk reduction. However, in the beginning this exercise will not produce valid estimates and is primarily intended to improve the method itself. Throughout the application a better understanding of expert disagreement will arise, which improves x-risk assessment and strategy directly. Only after much iteration would the method produce robust estimates that can inform policy. However, it is plausible that during the process a different path will be taken in which the quantitative approach becomes deprioritized in favour of a qualitative approach. The next step is to explore different methodologies. Some possibilities include

- a ‘breakout session’ at the 2020 conference based on the IDEA protocol
- a survey among the attendees of the 2020 CSER conference
- a number of workshops with different expert groups

- one-on-one interviews
- developing software to assist in the elicitation of beliefs
- facilitating crux elicitation of a pair of experts.

I have conducted a single pilot workshop. The next steps would be one to three semi-structured expert interviews, and collecting feedback on the project. The first steps will be taken by me (with feedback and supervision), and the bigger steps will probably involve more collaboration.

Literature

Used

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., ... & Sotala, K. (2019). Long-term trajectories of human civilization. *Foresight*, *21*(1), 53-83.

Baum, S. D. (2018). Uncertain human consequences in asteroid risk analysis and the global catastrophe threshold. *Natural Hazards*, *94*(2), 759-775.

Baum, Seth, Robert de Neufville, and Anthony Barrett. "A Model for the Probability of Nuclear War." *Global Catastrophic Risk Institute Working Paper*, no. 18–1 (2018). https://doi.org/10.2139/ssrn.3137081.

Bolger, F., & Rowe, G. (2014). Delphi: somewhere between Scylla and Charybdis?. *Proceedings of the National Academy of Sciences*, *111*(41), E4284-E4284.

Diaconeasa, M.A., Stewart, T. , Mosleh, A. & B. John Garrick (2018). A Quantitative Risk Assessment Study of the Initiation of an Inadvertent Nuclear War. *Proceedings of the One Day Workshop on Quantifying Global Catastrophic Risks*, B. John Garrick (Ed). The B. John Garrick Institute for the Risk Sciences, University of California Los Angeles, Dec 2018.

Hanea, A. M., McBride, M. F., Burgman, M. A., Wintle, B. C., Fidler, F., Flander, L., Twardy, C.R., Manning, B. & Mascaro, S. (2017). I nvestigate D iscuss E stimate A ggregate for structured expert judgement. *International journal of forecasting*, *33*(1), 267-279.

Hemming, V., Burgman, M. A., Hanea, A. M., McBride, M. F., & Wintle, B. C. (2018). A practical guide to structured expert elicitation using the IDEA protocol. *Methods in Ecology and Evolution*, *9*(1), 169-180.

Kuhlemann, K. (2019), "Complexity, creeping normalcy and conceit: sexy and unsexy catastrophic risks", *Foresight*, Vol. 21 No. 1, pp. 35-52. https://doi.org/10.1108/FS-05-2018-0047

Liu, H. Y., Lauta, K. C., & Maas, M. M. (2018). Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, *102*, 6-19.

Rowe, T., Beard, S., & J. Fox (manuscript). An Analysis and Evaluation of Methods Currently Used to Quantify Existential Risk.

Sandberg, A., & Bostrom, N. (2008). Global catastrophic risks survey. *Civil Wars*, *98*(30), 4.

Tonn, B., & Stiefel, D. (2013). Evaluating methods for estimating existential risks. *Risk Analysis*, *33*(10), 1772-1787.

To read

Many of these references are mere starting points into topics or fields of literature. I expect some to be dead ends, while others leading to fruitful methods and insights.

Avin, Shahar, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland, and Martin J. Rees. "Classifying Global Catastrophic Risks." *Futures*, February 23, 2018. https://doi.org/10.1016/j.futures.2018.02.001.

Barrett, Anthony Michael. "Value of Global Catastrophic Risk (GCR) Information: Cost-Effectiveness-Based Approach for GCR Reduction." *Decision Analysis* 14, no. 3 (August 24, 2017): 187–203. https://doi.org/10.1287/deca.2017.0350.

Cranshaw, J., & Kittur, A. (2011, May). The polymath project: lessons from a successful online collaboration in mathematics. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1865-1874). ACM.

Dafoe, A. (2018). AI governance: A research agenda. *Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK.*

Farquhar, Sebastian, Owen Cotton-Barratt, and Andrew Snyder-Beattie. "Pricing Externalities to Balance Public Risks and Benefits of Research." *Health Security* 15, no. 4 (August 2017): 401–8. https://doi.org/10.1089/hs.2016.0118.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. "When Will AI Exceed Human Performance? Evidence from AI Experts." *ArXiv:1705.08807 [Cs]*, May 24, 2017. http://arxiv.org/abs/1705.08807.

Gruetzemacher, R. (forthcoming). A Holistic Framework for Forecasting Transformative AI. *Big Data and Cognitive Computing*.

Hsia, P., Samuel, J., Gao, J., Kung, D., Toyoshima, Y., & Chen, C. (1994). Formal approach to scenario analysis. *IEEE Software*, *11*(2), 33-41.

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical assessment, research & evaluation*, *12*(10), 1-8.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis*, *10*(4), 305-326.

Lee, W. S., Grosh, D. L., Tillman, F. A., & Lie, C. H. (1985). Fault Tree Analysis, Methods, and Applications: A Review. *IEEE transactions on reliability*, *34*(3), 194-203.

Ord, T., Hillerbrand, R., & Sandberg, A. (2010). Probing the improbable: methodological challenges for risks with low probabilities and high stakes. *Journal of Risk Research*, *13*(2), 191-205.

Proceedings of the One Day Workshop on Quantifying Global Catastrophic Risks, B. John Garrick (Ed), Published by The B. John Garrick Institute for the Risk Sciences, University of California Los Angeles, Dec 2018

Tanaka, H., Fan, L. T., Lai, F. S., & Toguchi, K. (1983). Fault-tree analysis by fuzzy probability. *IEEE Transactions on reliability*, *32*(5), 453-457.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Appendix A: Questions for assessing the existential risk of a nuclear weapon exchange

[Test questions]

Probability of nuclear exchange
For these questions, assume no other catastrophic events happen.

1. What is your subjective probability that a nuclear exchange occurs that
   - does significant damage (at least 100 deaths or $10 mln. worth of damage)
   - has a maximum yield of 25 Mton TNT-equivalent (approximately 50 standard yield warheads)

in the next century (i.e. between January 1st, 2020 and December 31st, 2119)?

Please provide the lower and upper bounds, as well as your best point estimate.

2. What is your subjective probability that a nuclear exchange occurs that
   - does significant damage (at least 100 deaths or $10 mln. worth of damage)
   - has a yield of ***more*** than 25 Mton TNT-equivalent (approximately 50 standard yield warheads)

in the next century (i.e. between January 1st, 2020 and December 31st, 2119)?

Please provide the lower and upper bounds, as well as your best point estimate.

Residual risk
3. What is your subjective probability that **humanity undergoes a global societal collapse** in the next century without a nuclear exchange (i.e. between January 1st, 2020 and December 31st, 2119)?

Let's define global societal collapse as containing at least

- The worldwide loss of industrial activity within a decade
- Loss of Gross World Product of 50% or more within a decade
- 50% loss of world population or more within a decade
- Political disintegration: none of the current G20 countries are able to raise taxes within a decade

However, if 100% of humanity dies within one year, this does not classify as collapse, but as 'direct extinction'

Please provide the lower and upper bounds, as well as your best point estimate.

4. What is your subjective probability that **humanity undergoes direct extinction** in the next century without a nuclear exchange (i.e. between January 1st, 2020 and December 31st, 2119)?

Let's define 'direct extinction' as 100% of the human population dying within one year, and no descendants to humanity remaining alive.

Please provide the lower and upper bounds, as well as your best point estimate.

Impacts of a nuclear exchange

5. Given that a nuclear exchange occurs with a yield of **25 Mton TNT-equivalent or less**, please give your subjective probabilities of

- Collapse
- Direct extinction

Please provide the lower and upper bounds, as well as your best point estimate.

[Note: the probability of 'neither' will be calculated as follows:

$$p_{best}(\neg Col \ \& \ \neg Ext_{\ direct} | NX \leq 25) = 1 - p_{best}(Col | NX \leq 25) - p_{best}(Ext_{\ direct} | NX \leq 25)$$

$$p_{upper}(\neg Col \ \& \ \neg Ext_{\ direct} | NX \leq 25) = 1 - p_{lower}(Col | NX \leq 25) - p_{lower}(Ext_{\ direct} | NX \leq 25)$$

$$p_{lower}(\neg Col \ \& \ \neg Ext_{\ direct} | NX \leq 25) = 1 - p_{upper}(Col | NX \leq 25) - p_{lower}(Ext_{\ direct} | NX \leq 25)$$

]

6. Given that a nuclear exchange occurs with a yield of **more than 25 Mton TNT-equivalent**, please give your subjective probabilities of

- Collapse
- Direct extinction

Please provide the lower and upper bounds, as well as your best point estimate.

Impacts of non-collapse

7. Given that a nuclear exchange with a yield of **25 Mton TNT-equivalent or less** occurs without leading to collapse or direct extinction, how much existential risk would this post-exchange civilisation face in the remaining period after the exchange, compared the exchange not occurring?*

A factor of 0.5 means the society faces only half of the extinction risk we currently face. A factor of 2.0 means they face double the existential risk we currently face.

Please provide the lower and upper bounds, as well as your best point estimate.

* *For example, if a nuclear exchange occurs in the year 2070 that does not lead to collapse or extinction, how much different would the existential risk be in the period from 2070 to 2120 compared to the existential risk from 2070 to 2120 if no nuclear exchange had occurred?*

8. Given that a nuclear exchange with a yield of **more than 25 Mton TNT-equivalent** occurs without leading to collapse or direct extinction, how much existential risk would this post-exchange civilisation face in the remaining period after the exchange, compared the exchange not occurring?

A factor of 0.5 means the society faces only half of the extinction risk we currently face. A factor of 2.0 means they face double the existential risk we currently face.

Please provide the lower and upper bounds, as well as your best point estimate.

* *For example, if a nuclear exchange occurs in the year 2070 that does not lead to collapse or extinction, how much different would the existential risk be in the period from 2070 to 2120 compared to the existential risk from 2070 to 2120 if no nuclear exchange had occurred?*

Impact of collapse & recovery

9. What is your subjective probability that **humanity eventually recovers from collapse**?

Let's define recovery as:
Redeveloping a in industrial society with similar capabilities as humanity had globally around 1900.

Please provide the lower and upper bounds, as well as your best point estimate.

10. Given that humanity recovers, how much existential risk would this post-recovery society face compared to our current state (i.e. the existential risk from January 1st, 2020)?

A factor of 0.5 means the society faces only half of the extinction risk we currently face. A factor of 2.0 means they face double the existential risk we currently face.

Please provide the lower and upper bounds, as well as your best point estimate.