# A Conversation with Carl Shulman on September 25, 2013

**Participants:**

- Carl Shulman – Research Associate, Future of Humanity Institute
- Alexander Berger – Senior Research Analyst, GiveWell
- Sean Conley – Research Analyst, GiveWell

***Note:*** *This set of notes was compiled by GiveWell and gives an overview of the major points made by Carl Shulman.*

## Summary

GiveWell spoke with Carl Shulman about global catastrophic risks and existential risks. Conversation topics included the likelihood of such risks, the moral importance of future generations, and the possibility of encouraging government attention to these issues. Mr. Shulman also discussed people and organizations involved in this area.

## Global catastrophic risks and existential risks

The term global catastrophic risk (GCR) has been given multiple definitions. Some use the term very broadly to refer to problems such as financial crises and disasters that kill many people but only a small percentage of humanity. More stringent definitions focus on threats that could kill a large portion of humans or disrupt industrial civilization. Existential risks are catastrophes that end humanity's existence or have a drastic permanent disruptive effect on the future potential of human-derived civilization.

Existential risks potentially threaten hundreds of millions of years of human history in the future and as such are subject to special considerations. The effects of smaller disasters, or those from which humanity would recover, can be approximated in terms of lives lost, economic damage, and flow-through effects of those changes. Mitigation of such risks can be more readily compared to interventions such as bednets: both can be viewed as competing routes to increasing population, wealth, etc. Evaluation of existential risks depends more tightly on how one values future generations.

Some GCRs also pose existential risk, but most GCRs are much more likely to cause severe global harm than to permanently disrupt or destroy civilization.

## The importance of the future

A number of philosophers, including Peter Singer and Derek Parfit, have argued that a premature end to the human future would be very bad, because it is potentially

extremely populous and is likely to have a higher standard of living than the present and past.

There are others who argue that preventing future people from coming into existence is not negative.

There are some cases where benefits to future generations are invoked in support of incurring current costs. Examples include concern for long-term pension stability, supporting research, development, and investment, or policies involving immediate sacrifices to prevent future climate change.

However, it is rare for actors to systematically apply the view espoused by Singer and Parfit, the overwhelming importance of future generations, across policy issues. In debates about climate change some economists such as William Nordhaus have argued against placing high value on future generations, since this would seem to require that current people should also make large sacrifices in other areas, such as high saving and investment rates, and that these sacrifices are unacceptable to current people.

**People and organizations involved**

The community of people and organizations that explicitly identify as working on existential risk includes:

- Nick Bostrom and the organization that he runs, the Future of Humanity Institute (FHI).
- The Global Catastrophic Risk Institute (GCRI), founded by Seth Baum and colleagues. [The GCRI's research is not strictly limited to existential risk, but places strong weight on it.]
- The Cambridge Centre for the Study of Existential Risk, which is currently in development. It has a board of directors and one employee, and is currently applying for grants from academic funding bodies to hire full-time staff. Huw Price and Martin Rees are two of the co-founders.
- The Machine Intelligence Research Institute.
- The Institute for Ethics and Emerging Technology has an existential risk/emerging risk category.

Within the Effective Altruist movement, there are a number of people who put priority on the future but may not be doing direct work in this field. Mr. Shulman suggested that their lack of current engagement may be due to a lack of trusted recommendations for assisting the future, and that if GiveWell made a recommendation in the GCR/existential risk area, many people would likely be interested in contributing in this area.

In general, philanthropists in this area seem to be more focused on catastrophic risks than existential ones. There doesn't seem to be a major philanthropist putting

heavy weight on future generations and giving in the GCR area with that perspective.

Some other individuals and organizations that are working on or have worked on some aspects of GCRs, but not typically existential risk, include:

- Richard Posner, a legal scholar and judge who has written on catastrophes.
- Martin Weitzman, a climate economist, has written about extreme right tail climate risks.
- Within the climate literature, a number of people work on what the appropriate discount rate ought to be.
- John Broome, a philosopher at Oxford, has written about utilitarianism, population ethics, and climate change. In general, Oxford has a number of utilitarian-minded philosophers who write about the importance of the future, including at the Uehiro Centre for Practical Ethics.
- The Long Now Foundation hasn't done much work on GCRs or existential risk to date, but likely has an interest in the area.
- World Economic Forum's Global Risk section has done some research and provides examples of past work on GCRs.
- The Skoll Global Threats Fund has some interest in these issues.
- The venture capital firm Kleiner Perkins started a $200 million fund to invest in bio-defense and preventing pandemics in 2006.
- Warren Buffet has donated some money to the Nuclear Threat Initiative.
- Jason Gaverick Matheny, a Program Manager at the Intelligence Advanced Research Projects Activity and a former scholar at the Future of Humanity Institute, has written about a number of catastrophic risks.
- The *Bulletin of the Atomic Scientists* is concerned with nuclear GCRs and existential risk.

**Level and trajectory of existential risk**

The value of existential risk reduction depends on the trajectory of risk. If the level of risk per period is constant over time, then the chance of civilization surviving will decay exponentially with time. The Stern Review (published by the British Government in 2006) of the economics of climate change used an assumption that every year humanity has a 1 in 1000 chance of going extinct. This amounts to a 0.1% discount rate, or a mean lifetime for civilization of about 1000 years. While the Stern Review still found that future generations total welfare would be many times greater than that of the present generation, this pessimism results in a drastically lower estimate of future welfare than one which allows for rates of annual risk to fall.

There are a number of reasons to expect the risk of extinction could fall over time. First, many such risks are associated with technological transitions, and as time passes and humanity approaches the limit of potential technologies there will be fewer technological surprises. Second, as Steven Pinker has argued, violence has

declined and large-scale cooperation has increased over human history, and that trend may continue, which would reduce existential risk. Third, space colonization, which would separate humans by vast distances, would make it more difficult for a stochastic process to destroy the entire civilization. Finally, technological advances could create new ways to prevent disruptions. For example, surveillance or lie detection could improve, making it more difficult for rogue actors or mutual distrust to produce catastrophes.

Mr. Shulman expects Earth-derived civilization to still exist in a million years.

**Attempts to estimate the likelihood of human extinction**

Some works that discuss the likelihood of human extinction include:

- *Global Catastrophic Risks* edited by Nick Bostrom and Milan M. Cirkovic is a compilation of articles about candidate risks and general methods, with discussion of the evidence on the risks. Its bibliography is a good source of other references.
- A survey conducted of attendees of a conference related to this book asked for estimates of risks, and is available on the Future of Humanity Institute's website: http://www.fhi.ox.ac.uk/gcr-report.pdf
- *Our Final Century* (published in US as *Our Final Hour)* by Martin Rees discusses a wide range of issues with brief estimations.

While there aren't many estimates of extinction risk in aggregate, there have been estimates produced for specific risks. A larger number of estimations have been produced of catastrophic risks, some of which could become existential. These estimates include:

- Martin Hellman has written an article estimating the risk of nuclear war, although he does not discuss the consequences.
- A paper presented at a recent Philosophy & Theory of Artificial Intelligence (PT-AI) conference included surveys with predictions of the timeline and consequences of artificial intelligence (AI) development. Respondents included some of the most cited AI authors, members of an AI professional organization, and attendees of conferences on AI and the future of AI.
- The Intergovernmental Panel on Climate Change has estimated the probability of forms of climate change that would directly cause human extinction (such as Earth becoming like Venus), and found that to be very improbable. Less severe climate change might leave the planet habitable but cause severe damage that interacts with other problems such to cause collapse of civilization. Martin Weitzman has argued that the probability of severe damage from climate change should be higher due to model

uncertainty[1]: ultra-low risk estimates are often due to strong assumptions in a model which are insufficient to support strong predictions. FHI scholars have written about this issue[2].

A number of sources have talked about the evidence for catastrophic risks without committing to a probability.

- *Catastrophe* by Richard Posner
- *Worst-Case Scenarios* by Cass Sunstein includes discussion of precedents from past efforts, such as controlling the ozone hole
- Papers on the Future of Humanity Institute's website and on nickbostrom.com
- *The Concept of Existential Risk*, a paper by Nick Bostrom, is a good overview of the conceptual issues and has many useful references

Note that there is a selection effect on many published predictions of existential risk: those who believe the risks are larger are more likely to write about them.

Many estimates of GCRs do not include an estimate of the risk of extinction, given that a catastrophe occurred. For example, there are estimates of the risk of an asteroid impact, but rarely is the risk of extinction from that asteroid impact estimated.

**AI risk**

AI risk seems to pose a larger existential risk than the existential risk components of other GCRs, since conditional on other GCRs such as nuclear war the probability of outright extinction or permanent collapse still appears low. AI gets less attention than more immediate nuclear and bioterror threats, although there is some attention. For example, Richard Posner recommends additional research on the topic, arguing that it is potentially severe, but distant, so that only research should be undertaken now until the situation becomes clearer and/or the technology is closer. Compared to other risks, AI risk is particularly likely to become existential should an AI catastrophe occur, since the capabilities of problematic AI would increase over time, whereas the effects of GCRs such as nuclear winter would become less deadly over time and allow recovery.

**Potential for lobbying government**

---

[1] http://scholar.harvard.edu/weitzman/publications/modeling-and-interpreting-economics-catastrophic-climate-change

[2] http://www.fhi.ox.ac.uk/probing-the-improbable.pdf

Options for lobbying governmental organizations or national defense establishments to focus on GCRs vary heavily with specific risks. For example, encouraging government spending was highly successful for asteroids, an issue where it was productive to simply scale up funding for known methods, and there was no risk to increased efforts, because bad policy does not increase risk from asteroids. However, other areas are more complex, such as nuclear weapons. Funding levels are much higher, with billions of dollars spent every year by governments, and countries will take extreme steps such as sanctions or war to prevent nuclear proliferation. The existence of substantial current efforts means that spending may suffer from diminishing returns, and some of the policy options under consideration may backfire and actually increase nuclear risk.

One notable government-sponsored project was the Intelligence Advanced Research Projects Activity's prediction tournament, which was intended to improve the capacity of the intelligence community to anticipate geopolitical events. Philip Tetlock and other competitors have had reasonable success at predicting such events. One of the competing teams is working on using the model for technology prediction. The techniques could feasibly be used for better estimating GCR and existential risks.

**Existing funding**

- The Future of Humanity Institute's budget, stemming primarily from grants, the James Martin School, and some private donations, is currently $1.1 MM, including all overhead charged by Oxford University, conferences, and other programs. There are 12 full time employee equivalents, plus research associates. FHI receives academic grants reliably enough to continue to exist, but the grants often require that work be moved in the direction of funders' interests relative to its core research agenda. Such grants allow FHI to work on useful projects, but not as useful as they might otherwise do with more unrestricted funding, and the work of the Institute is also impacted by the time and human cost of repeatedly applying for grants.
- The Cambridge Centre for the Study of Existential Risk has 1 employee shared with FHI and other possible employees ready for when it gets funding.