# Reducing Catastrophic Risks: A Practical Introduction

Ryan Carey

September 10, 2015

So you're interested in playing a part in improving the future prospects for humanity? Welcome — lots of us are, but where are we to start? The risks are many, and we have only so much time to work out how to address them. Hopefully, this document will help by presenting an overview of the thinking done on this question.

It's in five parts:

1. Which risks should we worry about?
2. Approaches to mitigating them
3. What you can personally do
4. How does this differ from past advice?
5. Which organisations can I team up with?

## 1 Which risks should we worry about?

Step one for reducing risks to humanity is thinking about which risks to focus on.

The first key idea here, which is quite obvious once it's been said, is that the most important risks have to be those that are manmade because manmade risks operate on a shorter timescale. Life has existed on Earth for about 4.6 billion years and over that time it has withstood many natural catastrophes. Humanity has been around for a few hundred thousand of those but only for the last few thousand years have we performed tasks like writing, irrigation and mass industrial agricultural activities. The industrial revolution was only a few centuries ago. Since that time, many new

and transformative technologies have been developed, which we have only co-existed with for a short scope of our time on this planet which are radically impacting our collective global community. So the risks that are most important for the present century most all arise from our modern technologies.

This narrows the field considerably, but we're still left with lots of important possibilities. How big are the risks from climate change? From pandemics? Or from artificial intelligence (if that's something you worry about)? How should we prioritise and quantify these risks and go about addressing them?

Which do we care about? Well, we care about the risks that are serious but there's a threshold of seriousness that's especially important: we care most about the risks that are irrevocable. The unintuitive aspect is that if some disaster arises that sets technological development back by some decades, that's inconsequential when looked at on a geological timescale. From that long-run perspective we know that the Earth has a life expectancy on the order of at least hundreds of millions of years. Given how the size to which our civilisation could grow, and the number of years we could live, our main job, here and now, is to simply to make it through the next hundred.

How else might we choose risks to focus on? Well, we want to pick risks that can be reduced, and where the efforts to mitigate them are uncrowded so that we can personally make a bigger difference.

For the purpose of this brief strategic overview, it's not necessary to go over all of the risks, but it will be handy to have examples of some of the most pressing. The Future of Humanity Institute are the top investigators of existential (i.e. irrevocable) risks, and here is a fairly typical popular account of the risks that rank as the most concerning ones:

1. Nuclear war
2. Bioengineered pandemic
3. Superintelligence
4. Nanotechnology
5. Unknown unknowns

What's worth noting about these risks is that they're fairly diverse and predominantly technological in nature. Although there are varying opinions about which of these risks is the biggest, people don't usually estimate their top risk at more than 10x more probable than the next one. So since most of these risks are substantial and require specific technical expertise to understand, we will have to perform some

division of labour. It will be necessary for domain experts to learn about some particular risks, and for other nonexperts to perform risk-management in policy and other contexts.

What some people find surprising is that climate change did not make the list. It would probably make the top ten but even if it did, it is more crowded — there is already an established community of researchers and advocates working on that problem, which cannot be said for these five. Another thing that bothers people is the mysterious point 5. If unknown unknowns are on our list — scenarios that can't even be predicted by our best experts — then what are we supposed to do? Perhaps we must find better experts or build some attributes of general resilience. Regardless, the fact that the risks are not only diverse but also sometimes obscure favours responding to multiple risks at once using more general strategies.

So having surveying the kinds of risks involved, it's time to get closer to the practical question of how we can fit the reduction of these risks into our lives... But we're not quite there yet. There's a bit more strategy — what are the kinds of activities that people would be doing that would make us safer from technological risks? This is the focus of the next section.

# 2 Approaches to mitigating tech risk

## Research

We have said that the big risks are technological and many are still being discovered. So it's worth investing some resources into better forecasting emerging technologies, though this is a notoriously difficult task. It is also valuable to build safety features into these technologies, ideally before they are implemented (e.g kill switches for strong A.I (the ethics of non-biological euthanasia momentarily aside) and to reach out to others who will do the same. Hence our first three approaches to tech risk mitigation:

1) **Forecasting and macrostrategy.** e.g. The Future of Humanity Institute, The Centre for the Study of Existential Risk, The Open Philanthropy Project, Forecasting Science and Technology (ForeST), AI Impacts, technology assessment institutions, The Institute for Future Studies.

2) **Tech safety engineering.** e.g. The Machine Intelligence Research Institute.

3) **Academic outreach.** e.g. Stuart Russell, The Future of Life Institute, The Global Catastrophic Risk Institute, Bulletin of Atomic Scientists, Edge event attendees, employees of Google Brain or DeepMind.

## Security

A second critical component of the growing tech risk–mitigation infrastructure will be having professionals in the intelligence and cybersecurity communities who can liaise between researchers and government. There is also a role for more eclectic research into how to rebuild if something goes wrong.

4) **Cybersecurity & International security (± outreach).** e.g. Bruce Schneier, Palantir, David Denkenberger, safety barriers, surveillance generally (perhaps).

## Policy making

Another critical link in the chain from research to implementation will be the policy makers. Currently, the number of people who have or are pursuing a theoretical understanding of risky technologies is more than the number who are willing or able to implement these in government. Currently, we don't know what policies to promote to reduce catastrophic risks, so there will be a slow transition through policy development and eventually to implementation.

5) **Tech policy development (± outreach).** e.g. The Centre for the Study of Existential Risk, some The Future of Life Institute grantees, The Center for International Security and Cooperation, The Royal United Services Institute, The British American Security Information Council.

6) **Politics.** Ministers of defence, foreign affairs and science.

7) **Public service.** In intelligence, cybersecurity, intelligence research and defence.

## Broad-based outreach

For some individuals, the best way to reach individuals who can help with the above problems is going to be to do broad-based community outreach. The presence of such organisations gives a plausible story for how the above activities were supported and coordinated.

8) **Future-building outreach.** e.g. The Centre for Effective Altruism, The Future of Life Institute, The Center for Applied Rationality, The Machine Intelligence Research Institute, The Centre for the Study of Existential Risk, Stephen Hawking.

### Funding

Lastly, many of the above projects, especially the broad-based outreach but also some of the forecasting and safety engineering efforts will require philanthropic contributions.

9) **Funding.** e.g. Elon Musk, Jaan Tallin, Skoll Threats Fund, Good Ventures, tech-related foundation program managers, partners for venture capitalists, effective altruists.

This list is not exhaustive, other activities could be included, for example a special category could be made for technological journalism or for strategic investment in risk-relevant tech companies but most of the major bases are covered.

## 3 What can I personally do?

Finally, let's consider concrete personal plans, with an emphasis on careers, which is where we spend the majority of our efforts.

For someone who cares about tech risks and is already working in a relevant domain, such as an elected politician, probably the best thing for them to do is to stay at it or just pivot into the nearest of these risk mitigation approaches. For researchers in relevant fields that are not specific to any one tech (think economics, or some parts of philosophy), a useful approach will be to try and collaborate with other risk-focussed organisations.

For people already working in less-relevant fields there are often still opportunities to cross-over to discussing tech. For example, Dylan Matthews, a journalist at Vox is making a career of discussing these important issues. Careers like marketing, management and executive assisting work can fit into basically any of these pathways, depending on the job opportunities that arise for a given individual. Clearly, not even in a research organisation should every single individual be a complicated genius. Plenty of more rounded individuals are needed everywhere to make ambitious projects run properly and interact smoothly with other organisations.

For people who are just starting out, it's going to be an issue of finding their comparative advantages. Policy-minded people should be seriously advancing that interest to fill out the movement with a skillset in which it is currently lacking. Obviously skills in networking and coalition-building are required. People with a talent for theoretical computer science and discrete mathematics should see whether they can usefully contribute to The Machine Intelligence Research Institute. People with great academic and technical ability may suit academic research, and so on. Funding is another task, like management and executive assisting, that nearly anyone can contribute to. Though probably less important than aptitude, personal interest in a subject is also a factor.

There's a lot more discussion to be had on the topic of career selection, but much of it depends on the person, so the best I can hope is that this framework will guide that ongoing discussion.

# 4 How does this differ from past advice?

There's been articles about risk reduction before but some are focussed on specific risks or explain detailed results of investigations. Here, I've tried to give a readable strategic overview.

So how does this strategic overview differ from what someone else might write, or what someone might have written several years ago?

For good reasons, over the past decade, risk researchers have focussed on building a solid theoretical basis for risk assessment. Now that our perceptions of the top risks are stabilising, the emphasis is moving toward outreaching to academics and policy makers who can advise and drive the implementation of these policies. In that regard, the list is different from what it might have been a few years ago.

Another way my thinking differs from what's come previously is that I try to take replaceability seriously: being involved in developing risky technologies like AI or surveillance might not be a terribly bad thing, if it is going to be done anyway, because then you get to sound alarm bells if risk becomes elevated, while networking with people in an important space. The risk associated with speeding up a technology's development might be outweighed by the networking benefits from participating in that field.

On a related note, discussions are taking a more conciliatory tone towards technology developers than previously. This is because there are more discussions with both technologists and policy makers than there were. In the first party, there is some

growing concern about scaremongering including distress regarding numerous articles covered with Terminator pictures. Although these images irk AI risk proponents similarly, they are understandably received as antagonistic. In policy makers, there is a desire for unified answers from technologists, security experts and risk researchers. So building bridges between the above is critical for moving the discourse to the next level and if only for this reason and no other, all of these are plausible areas of work to aim to get into.

Since the outreach is supposed to be more targeted to people in tech and security, and oriented toward action, I've put 'future building' as a placeholder for the kind of movement that one would want to develop, though the rationality and effective altruism communities are the main ones that have so far been working in and around this space.

Apart from these points, I think I've just summarised what a lot of people have for many years been thinking.

# 5 Which organisations can I team up with?

It's very useful to learn about organisations working on addressing these problems, to assess what one might personally be able to contribute.

Of the organisations stated above, a handful deserve particular mention:

- The Machine Intelligence Research Institute (MIRI): performs technical AI safety research using discrete math
- The Centre for the Study of Existential Risk (CSER): applies various sciences to risk forecasting and reduction
- The Future of Humanity Institute (FHI): leads in macrostrategy research, founded by Nick Bostrom.
- The Future of Life Institute (FLI): fundraises and does academic outreach. Distributed funds from Elon Musk and hosted a successful AI safety conference in Puerto Rico.
- The Global Catastrophic Risk Institute (GCRI): research and academic outreach in relation to resilience and other topics.
- The Open Philanthropy Project (Open Phil): an offshoot of GiveWell that allocates funds on behalf of its partner foundation Good Ventures for catastrophic risk reduction and other causes.

Over the next decade hopefully many more such organisations will be founded.

To learn more, you can look up these organisations, and you can also read more about all of these topics at www.existential-risk.org.

# 6 Conclusion

The kinds of people we need to reduce tech risks are different from what we needed a decade ago. We need to act in at least nine domains: 1. Forecasting and macrostrategy. 2. Tech safety engineering. 3. Academic outreach. 4. Cybersecurity + International security 5. Tech policy development. 6. Politics. 7. Public service. 8. Future-building outreach. and 9. Funding.

*Thanks to Owen Cotton-Barratt, Niel Bowerman and Haydn Belfield for feedback on an earlier draft.*