Effective Altruism
Geneva

Strategic support for the x-risk mitigation ecosystem

Florent Berthet - Oliver Bramford - Konrad Seifert

January 2018

SUMMARY

In this white paper we explain what we view as the most pressing issues of our time, and how we plan on dealing with them. We do a quick overview of the current situation and detail in what ways our approach could have great positive impacts. We wrote this document so that people in the EA community could know about our project, give feedback and get involved. Here's our mission statement:

Mission

To improve the odds of favorable long-term outcomes for civilization mainly by reducing existential risks.

Plan

1. **X-risk strategy needs assessment**
2. **Get the most accurate picture of x-risks**
3. **Find the best opportunities for action**
4. **Execute on these opportunities**
5. **Increase the number of talented people working on x-risks**
6. **Increase the effectiveness of people working on x-risks**

INTRODUCTION AND PROBLEM

We are either at the beginning or at the end of our civilization. Therefore, if we agree that the far future may hold tremendous value — as is often argued within the effective altruism community — then we should create the most favorable conditions for our civilization to survive its technological adolescence and to thrive in the long run.

More and more organisations have started working on this mission, but they are facing serious challenges, among which:

- There is no clearly defined and up-to-date landscape of x-risks.
- It is hard to know what the best approach is and how current efforts complement each other.
- As a result, collaboration between existing organisations has a lot of room for improvement.
- Organisations have trouble finding qualified talents.
- They also struggle to translate their research into concrete action.
- Funding is still not on par with the scale of the problem.

Fortunately, the global community has been making great progress in the last few years, leading to important safety developments. As an example, concerns about the risks of AI, spearheaded by the Machine Intelligence Research Institute and the Future of Humanity Institute, have led to a worldwide awareness and recognition of the problem. These concerns, which were frequently considered weird or absurd just a few years ago, are now shared by highly respected people such as Bill Gates, Elon Musk, and the late Stephen Hawking. An increasing number of AI researchers also agree on the importance of AI technical safety, and tech giants such as Google, Facebook, Apple, Amazon and Microsoft have joined the Partnership on AI, which aims at preventing a misuse of this powerful technology.

As a side note, this is a good illustration of the influence that a few dedicated people can have on the world.

But AI is just an example. Our civilization is facing many more threats, and while it is true that an increasing number of people are now working on AI safety (though probably still not enough), the same cannot be said for most other potential sources of existential risks, which include biotechnology, nanotechnology, global pandemics, nuclear conflicts, autonomous weapons, natural threats, etc.

The odds of a major catastrophic event may vary by several orders of magnitude between these risks. We still have a limited understanding of this whole topic, and it is hard to put numbers on them, sometimes even approximate ones. But the stakes for humanity are so high that staying in this state of ignorance is very dangerous, especially since technology continues to advance and its risks are rapidly growing. It is therefore urgent that we allocate more resources on trying to get a more accurate picture of these risks and how to mitigate them.

SOLUTION

We cannot answer to all the challenges our community is facing, but we have good options to deal with some of the major ones. As stated in the summary, we will focus on the following objectives:

1. X-risk strategy needs assessment

Survey and interview key stakeholders in the x-risk community to clarify the most essential needs and opportunities. This growing ecosystem will face ever-changing challenges, and it is crucial to gather constant feedback in order to offer as much value as possible.

2. Get the most accurate picture of x-risks

By reviewing the literature and talking with the most knowledgeable people in relevant fields, we will gather the information we need to have the best possible view of the current landscape of x-risks and the people working on them. We will then share our findings with everyone. The value for other organisations will be to know who is working on what, to see whether their work is too much redundant with another team’s work, and to find partners or experts who could give them useful information.

In the medium-term, we may want to create an x-risks mapping and quantitative modelling tool. This would allow any user to see the potential causes of various x-risks, and what influences these causes. Anyone would be able to put his own estimates and confidence intervals on any element and see how that changes the whole picture. Each user would also be able to see what the other users' estimates are, which would be useful to notice where people disagree.

3. Find the best opportunities for action

By comparing our updated picture of the ecosystem with our list of x-risks mitigation opportunities, we will find gaps to fill. The lowest-hanging fruits will be our priority. A good analogy is the work done by Charity Science Health: instead of trying to do more of what was currently being done, they scanned what was not being done and identified a gap with a promising solution, which was to send text messages to remind indian families to have their kids vaccinated. Our goal is to replicate this approach, but for projects focused on improving the long-term future of humanity.

4. Execute on these opportunities

For each gap we find, we will see whether an existing organisation would be better positioned to fill it or if we should deal with it ourselves. Potential courses of action may include:

- **Political advocacy to promote and enforce safety measures** (AI safety, biosafety, nuclear safety, etc.). Geneva may very well be the best place for such an initiative. Here is a quote from Genève internationale: "Home to 99 international organisations, 250 NGOs and 173 State representations, Geneva is at the heart of international cooperation. No other place in the world hosts such an important number of key global players as Geneva. From public health to humanitarian affairs, migration or protection of the environment, they address today's most pressing challenges."

- **Research on specific risks to find new promising solutions.** Given how little research there has been on finding new ways to reduce x-risks, it wouldn't be surprising if the lowest hanging fruits were still hiding in plain sight. This could prove very cost-effective to work on, and if not, we should find out rather quickly.

- **Research to find risks that haven't been noticed yet (unknown unknowns).** For the same reason as above, there might be risks that we have never thought about before. It will be harder to work on this since it requires us to think much broader than when trying to find solutions to a given risk, but it might also be worth a try.

- **Development of products or services to reduce specific risks, or grantmaking for such ventures** (e.g. a software that can scan any AI-related source code to detect dangerous goal structures; a guide for world leaders on how to best react in case of a pandemic; a defensive technology to counter autonomous weapons, etc.). This project could eventually require hundreds of people. And since it is easy to come up with ideas that are likely net positive, the question will be whether each idea will be cost-effective compared to the best giving opportunities.

5. Increase the number of talents working on the long-term future

Since a lot of organisations struggle to find good talents, it is crucial that more talented people hear about x-risks, and that the ones who already know about these topics actually want to work on them. We see three ways to do that:

- Fundraise for x-risk related projects (in Geneva and globally). In Switzerland alone, foundations move between 1 and 2 billion dollars every year. We aspire to move some fraction of that money to promising projects.

- Build a physical hub in which x-risk focused people would love to live and work. This hub is an ambitious project on which we have already made progress. It deserves its own white paper, which we will deliver if and when we decide that the project is worth pursuing.

- Spread x-risks awareness through events such as conferences and workshops (some of which could be done in the hub). EA Geneva has been having good results by doing these events, and will continue to do so in the future.

6. Increase the effectiveness of people working on the long-term future

- Direct talents toward the most promising work, according to their personal fit and potential impact. By knowing which gaps should be bridged and by knowing the needs of other organisations, we will be able to advise people who would like to work on impactful projects. We will likely work closely with other actors in the field, such as 80000 Hours.

- Give them the tools, training and support to make them as productive as they can. This can be done online or on-site. The hub could be a great place for such training and support.

- Surround them with a fantastic community, full of passionate and brilliant people who want to improve themselves and the world (here again, the hub could be very instrumental).

- Improve other x-risk organisations' impact by helping them go beyond the research phase and actually implement solutions based on their findings. Work involving complexity science is currently being done at EA Geneva by Max Stauffer toward that very goal.

TEAM

Oliver Bramford: currently volunteers at EA Geneva by co-leading the meta-policymaking research project, and works as a digital marketing consultant. Oliver is a Philosophy graduate, used to work for Impact Hub, and has ten years' professional experience supporting diverse startups and small businesses.

Konrad Seifert: Co-founder and Executive Director of Effective Altruism Geneva, working to foster the use of reason- and evidence-based methods. Konrad studied two years of international relations and dropped out to launch EA Geneva. He is now pursuing an MSc in Data, Economics and Development Policy on the side.

Florent Berthet: co-founded Altruisme Efficace France. Founded a democratic school in Lyon. Lives in a cohousing place in which he has been involved for 3 years. Teaches entrepreneurship to engineering students. Did the French translations of Sam Altman's *Startup Playbook* and Luke Muehlhauser's *Facing the Intelligence Explosion*.

CONCLUSION

We have several promising opportunities to improve the odds of long-term positive outcomes for our civilization. The current global efforts are hampered by preventable causes: lack of actionable information, limited access to talents and funds, limited political influence. Our mission is to bring concrete solutions to these issues, and we believe our plans have a good chance of working. Most importantly, we are confident that our team has everything it needs to push through any roadblock and to adjust its course as necessary.

HOW TO GET INVOLVED

There are several ways you can help with the project :

- **Give us feedback.** If you have suggestions or questions, please contact us (see below), we would love to know what you think.
- **Tell us about your current work and pain points.** A more accurate picture of the landscape will allow us to help you better.
- **Volunteer to work with us.** There is plenty to do.

- **Suggest a project.** We want to incubate research projects and highly promising work (products, services, political campaigns, etc.). Please let us know if you have an idea, even if you don't want to work on it yourself.
- **Help us financially.** Either by giving directly to the project through EA Geneva or by talking to donors who can do the same.
- **Grow our network.** If you know people who can join or help us in any way, please put us in touch.
- **Come live with us in our next share house.** Until we have a full-sized hub, some of us will live in share houses in Geneva. Fill out this form if you are interested in joining us.

ADDITIONAL RESOURCES

- A visual presentation of how a hub could look like
- Some projects we could incubate

CONTACT

- Florent: florent.berthet@gmail.com
- Oliver: oliverbramford@gmail.com
- Konrad: konrad@eageneva.org