[Existential Risk / Opportunity] Singularity Management

Oct 2016

Contents:

Alexei Turchin's Charts of Existential Risk/Opportunity Topics

by James Blodgett

We are interested in this publication in what I call existential risk/opportunity singularities. A lot has been written about various aspects of this. Someone should summarize this material. Someone already has.

Alexei Valerievich Turchin is a transhumanist and a member of the Lifeboat Foundation Futurist Advisory Board. He has published several relevant papers and two books. He has developed several charts that diagram and summarize many ideas in many areas of existential risk and existential opportunity. They are a good introduction to thought and literature in those areas, and a map for further reading for those who want to explore in more depth.

I recommend Alexei's charts to readers as a way to gain an overview of many topics. They seem to be good summaries of areas with which I am familiar. They taught me a few things about areas with which I was not familiar.

This is a list of areas covered in Alexei's charts. Each topic separated by semicolons has a separate chart.

Immortality
Immortality; Existing methods of life extension; Digital immortality; Quantum immortality

X-risks
Typology of x-risks; Typology of human extinction risks

Prevention plans
Existential risks prevention

Different risks
Nanotech risks; Biorisks; Map of nuclear risks

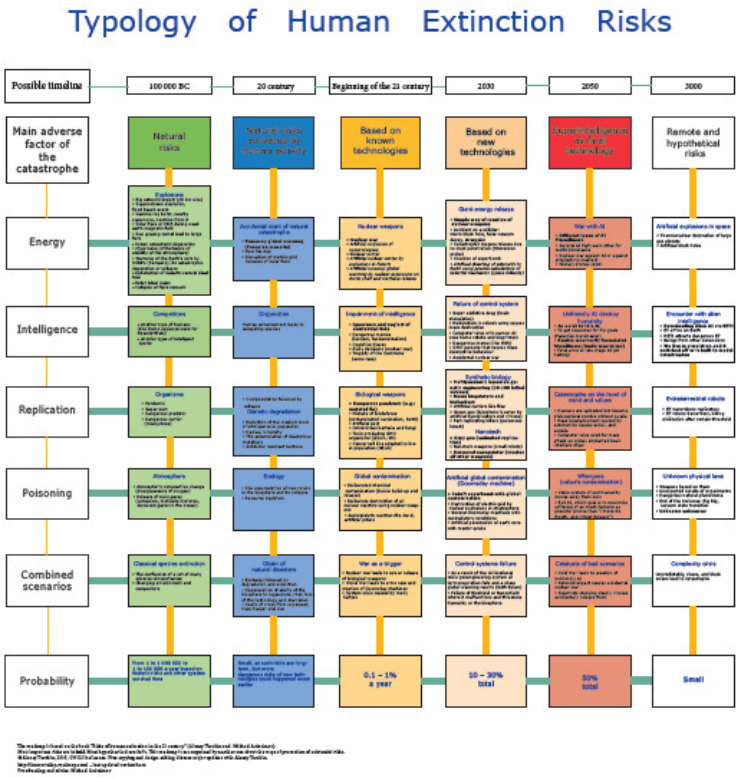
Structure of the global catastrophe: different approaches
Double scenarios of a global catastrophe; Causal structure of a global catastrophe

Levels of the global catastrophe
Levels of degradation following a global catastrophe

AI safety
AGI failures mode and levels; AGI safety solutions

Probability
Simulation; Doomsday argument; How to survive the end of the universe

Below is a sample chart. Don't try to read the small print. Download PDFs of these charts by followng links at: http://immortality-roadmap.com/sample-page/ . This chart is the Less Wrong version.



Interview with Alexei Turchin

Interview of Alexei Turchin by James Blodgett

Blodgett: What are your goals in constructing your charts? What can our readers learn from them?

Turchin: You call them charts. I like to call them maps. I have several goals with maps. I want to solve many problems in this area: the problem of friendly AI, the problem of preventing x-risks, the problem of achieving immortality, and many others. We have many ideas about these things. It is time to bring order to these ideas. I hope that my maps will be not only an encyclopedia of existing ideas, but also an instrument for creating new ideas, for finding the best of them, and for promoting them when found. Many people have asked me if they can print my maps and put them on their walls. My English texts never had such success.

I hope with my maps to create new ways of thinking about these things which use the graphic processor in the human brain. My maps are attempts to calibrate this and make it happen.

Blodgett: You have posted frequently on the Less Wrong website, established by Eliezer Yudkowsky. I am somewhat ambivalent about him. His reasons for worrying about AI are brilliant. Bostrom quotes him extensively, and they plan a book together. However, I think that Yudkowsky has overpromised on the idea of being able to prove mathematically that an AI implementation is friendly. I think that the concept of friendliness is too diffuse to permit a mathematical proof, and I note that Yudkowsky has not announced one yet. Also, the use of Bayesian reasoning that he promotes is useful, but doesn't convey magic superpowers. Yudkowsky seems to want to be a superintelligent Harry Potter. He has written fan fiction that puts an avatar of himself in that role. However, the guy is obviously smart and has some good ideas.

One of my objectives for EROSM is to is to give readers a picture of our field and of ways they might contribute. Yudkowsky's Less Wrong website (at http://lesswrong.com) has become a platform for debate and publication in our field, and also for meetups. I have respect for some of the people who post there. Since I am not a member and I am only marginally knowledgeable about how Less Wrong is set up, I am not the best person to present it to readers. You sometimes post there. Would you be able to tell readers a bit about it? Could they contribute there?

Turchin: In general I agree with your opinion about EY. I don't believe in the possibility of a mathematical proof of friendliness, because we still don't have a precise definition of what friendliness is. But the decision theory work of EY and some of the Less Wrong people is useful and impressive.

I think that the approach of using a mathematical proof is too complex to be solved in a timely fashion. We should try some simpler approach to developing friendly AI, an approach which may be less reliable, but which can be implemented by most actors in the field. I wrote about one possible approach today. Ask me about it below and I will share it with your readers.

I post often on Less Wrong, but I am annoyed by their strong downvoting system which can result in strong confirmation bias.

Blodgett: Can you suggest other ways our readers might contribute?

Turchin: I always need help with grammar, editing, and links for my maps, as well as translation of some map into English from Russian.

I also need attention. I have developed my maps to contribute to thought in this area. This only works when thinkers in this area are able to learn from them. I invite your readers to see what they can learn from them.

All maps are linked at: http://immortality-roadmap.com/sample-page/

Blodgett: You mention recent thoughts about developing friendly AI.

Turchin: I have written an essay. Its title is: Dissolving the AI Control problem via personal ascending.

The simplest model of AI creation describes the interaction of two entities: the creator of the AI and the AI. A more complex model would use a systems approach and add other stakeholders, for example: society, other AI projects, programmers, the employer of those programmers (a company, or a state), future generations, future states of AI development, and perhaps even space aliens, if they exist.

Adding more stakeholders makes the problem of friendliness more complicated. For example, if I add "society," I have to add into the AI not only my values but also values of other people who are unknown to me, values that I might not share. If I add the existence of other AI projects, and if I consider that mine is more likely to be friendly, I might conclude that it is important that my project be the first project to be implemented so it can prevent the others from realization. This may contradict other of my values, as it may require violence.

It is clear that the more stakeholders I add, the more intractable the problem becomes. It seems worthwhile to consider an ideal situation where the aforementioned problem doesn't arise, an ideal situation with the smallest possible number of stakeholders. Let us assume that the project will have only one stakeholder and it will be me. In this case there is a straightforward solution to the control problem: use self-improvement to become a superintelligence myself.

The positives of this idea: there are no intrinsic risks for me and my value system. My value system will naturally evolve on its own logic, and I may control the rate of my ascending.

But there are still risks of mistakes and unpredictable consequences. Wireheading, losing the meaning of life, and memetic hazards are still here.

At the level of this model it is not necessary to consider risks to others because there are no others in this model. There is no problem of communication, and I hope I will not become a paper clip maximizer. Even if my values evolve in a strange way, I still will think of them as "my" values, and will be satisfied with them.

NB: This idea needs more rigorous evolution and it is nowhere proved as safe. It is just a good-looking idea. Also, the chance that I personally will reach superintelligence first using self-improvement is small because many other capable people would be using the same techniques and instruments, so others are likely to be first. The essence of the idea is to ascend a person or persons with good intentions, so it doesn't have to be me.

Now, as we have something that looks like a working solution, I will try to adapt it to the existence of other stakeholders. My values seem to be not bad for other people:

1) I do not want to become a serial killer, so I will override any tendency in that direction.

2) I am interested in other people so I will keep people alive.

3) I am interested in preventing death and suffering.

4) I will create Tool AIs to solve practical task like fighting other AI projects, but I will understand how they work, and prevent them from evolving

5) I have sufficient understanding of human ethics so that I will not do obviously bad things, and my understanding will naturally evolve.

6) I will control the rate of my improvement and thus prevent risks of too quick improvement.

7) "Me" here could be a group of people from the beginning, connected by shared values, effective social practices and neuroimplants, perhaps tied together in a group mind implemented by physical interconnections between brains.

8) Merging of minds and consciousnesses in one large experience and brain could also ascend large groups of people, and would help the problem that those who are not included might feel left out.

The main problem with this approach is technical. We don't have the technology to implement it yet. Development of methods of human self-improvement lag in comparison to progress in computers. However, technology is advancing, so we can at least hope for development of a technical solution, and it would be wise to have considered this possibility philosophically so we have some idea of how to proceed should that technology become ready.

Blodgett: This has got to be a first. We are publishing an article in the midst of an interview.

Improving oneself is an interesting solution to the AI control problem. However, I see three problems with the approach. One is that the technology for turning a human into a superintelligence does not yet exist, and may never exist. We can hand-wave that away, as you do, by stipulating that the technology might come into existence, and that we are considering that contingency. You are somewhat aware of the second problem, which is that your good intentions might not survive the transition. However, I agree that finding someone with firm good intensions and hoping that those good intentions survive a transition to superintelligence might be our best chance. To my taste the third problem is your assumption, an assumption that others discussing this topic sometimes also make, that an appropriate solution is fighting other AI projects for priority, and in a sense taking over the world before other AI projects are able to do so. It seems contradictory to me

that some assume that a "friendly" AI will want to shut down other projects. That doesn't seem friendly to me. Even you, in your paper, say that shutting down other projects may contradict your other values, as it may require violence.

However, I do see the point that, given that the AI is much more competent than any other actor, it may decide, and even we may decide, that the world is best run by the AI. If we are going to solve this, I think we need some way to allow some version of AIs running many things while still having humans retain some form of control that will be acceptable to the rest of us. This is a hard problem. However, it is in some ways similar to the problem that was solved fairly well by Enlightenment philosophers, whose philosophies gave birth to modern representative democracies. Perhaps we can stir AI into this mix.

An essay within an interview is an interesting form of publication. I think it belongs here, as a valid window into your thoughts. Socratic dialogs start like that, with a partially finished idea that is polished in the give and take between Socrates and those who discourse with him. You might continue to polish your ideas, but so might EROSM readers. (If readers make use of these ideas, cite Alexei and EROSM.) I think Alexei's ideas are a good starting point for further thinking about how to solve these problems, and hopefully an inspiration to readers to think about similar issues.

Alexei, thank you for sharing this with us.