Book review: Toby Ord: The Precipice - Existential Risk and the Future of Humanity

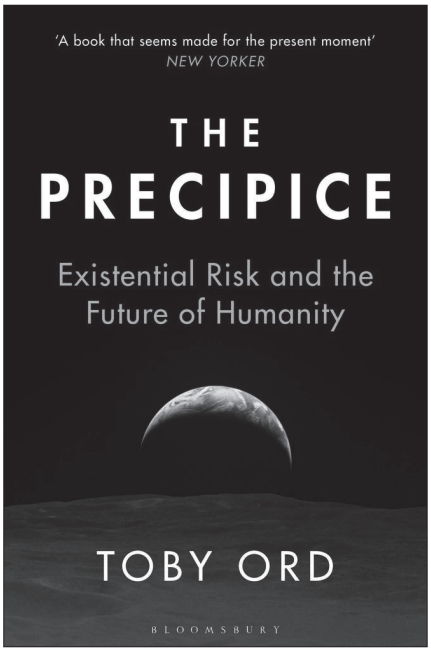
Soydan, Tolga

Veröffentlichungsversion / Published Version
Rezension / review

Empfohlene Zitierung / Suggested Citation:
Soydan, T. (2022). Book review: Toby Ord: The Precipice - Existential Risk and the Future of Humanity. [Review of the book *The Precipice: Existential Risk and the Future of Humanity*, by T. Ord]. *Intergenerational Justice Review*, *8*(1), 24-25. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-86395-3

Toby Ord:
The Precipice: Existential Risk and the Future of Humanity

Reviewed by Tolga Soydan



Imagine yourself rolling a dice, but instead of winning at a game, you find yourself rolling the dice on the fate of humanity, having a chance of 1 in 6 of destroying it over the next hundred years. Would you do it? Probably not, unless you are ridiculously confident or careless. In Toby Ord's book *The Precipice: Existential Risk and the Future of Humanity*, the senior researcher at the Future of Humanity Institute in Oxford argues that unless humanity does not take the possibility of existential risks more seriously, it stands the same chance of getting itself or its potential destroyed in the next century. So, the question arises: Why do we all roll the dice with such stakes?

Ord's ambition is clear: Showing humanity the risks it faces, warn us and even more, showing us the heights we could theoretically achieve in the long term, if we play our cards right.

To accomplish this, Ord divides *The Precipice* into three parts.

The first part, *The Stakes*, takes the reader to the humble beginnings of mankind, how we tamed nature and worked together and eventually reached the top of the food chain. But with great power also came great potential for destruction, reaching its practical pinnacle through the use of nuclear weapons in the Second World War. 1945 therefore marks the very beginning for Ord, where we reached the Precipice: the state in which humanity eventually possesses the means to destroy itself. Ord fears that there may be too much of a difference between our power and our wisdom to wield such power responsibly at the moment, putting us in a situation of higher existential risk.

Existential risks are defined as all the risks that could destroy humanity, leading to its extinction or permanently destroy its future potential, for example by getting locked into a dystopian scenario or suffering a permanent social collapse. Existential catastrophes are impossible to be undone and can never be allowed to happen. The importance of the matter is founded in the possibility of the trillions of people who could be born in the future, if we manage to avoid existential risks, as well as in all the lives before us that made the present possible. Ord calls the protection of existential risks an „intergenerational global public good" (59), as it especially benefits future humans. This good is insufficiently funded, comparing the billions of dollars that are spent on the work on AI to the millions of dollars that are spent on making sure AI is aligned with human values. One further glaring deficit in avoiding existential risks is the lack of a centrally coordinated institution.

In the second part, *The Risks*, Ord divides the existential risks into natural, anthropogenic and future risks. The natural risks are cases such as an asteroid or comet impact, supervulcanic eruptions or stellar explosions. He argues that we are well equipped in the case of a potential asteroid impact, because we have identified over 95% of the dangerous objects. As for stellar explosions and supervulcanic eruptions, the fossil record gives reasons to be fairly optimistic that those risks will stay minimal in the foreseeable future. Still, Ord pleads for more research on the field. Compared to the anthropogenic risks, he estimates the danger of natural existential risks a thousand times smaller (87).

Anthropogenic risks are risks such as nuclear weapons, climate change and general environmental damage. Even though each of those risks presents more of an existential risk by itself than all the three natural risks combined, Ord suggests it would be speculative to assume these anthropogenic risks to be sufficient to destroy humanity as a whole or its long-term potential. Nevertheless, Ord is in favour of more research on the effects of anthropogenic risks as well.

Ord finally locates the greatest danger for humanity in future risks connected to technology. He closely inspects the dangers of pandemics and biotechnology, unaligned artificial intelligence, dystopian scenarios and a few other risks, such as nanotechnology. Talking about pandemics, Ord highlights the dangers of biotechnology and information hazards, as unfiltered public information could lead bad actors to try and capitalize on the available technology and release deadly viruses. To date, the hypothesis that SARS-CoV-2 escaped from a laboratory in Wuhan, China, has not been completely dismissed. But Ord's main concern seems to be unaligned artificial intelligence, where he estimates the risk over the next hundred years to be on a 1 out of 10. If humanity were to successfully create a general AI smarter than human beings, our own fate would not necessarily be in our hands anymore. We do not know how to implement our values into AI, and yet we steadily upgrade the capabilities of AI making it more and more likely to put ourselves at risk.

In the third part, *The Path Forward*, Ord maps out in detail how he calculated the risks we could potentially face, how those risks could combine and how specific risk factors such as climate or economic failures could raise the danger of existential risks, and how specific safety measures could in turn lower it, such as achieving peace between the powerful nations. In addition, he urges us to re-evaluate the way we deal with risks on a theoretical and

practical level, strongly advising a more centrally organised policy making and binding powers to protect humanity and suggests representatives who stand in for future generations.

Looking to the future, Ord proposes three phases in which humanity could fulfil its potential. First, we have to reach Existential Security. For him this means to preserve and protect our potential by taking the risks seriously and managing them from their onset or avoiding them. The second phase, the Long Reflection, should be the time humanity literally spends time reflecting on the road it wants to take, choosing its best options. The third and last phase should see us achieving our potential. He keeps this section quite vague, explaining that humanity should first focus on reaching security.

As the state of knowledge on this field is quite young, he advises researchers to be more specific on possible risks and to be cautious about what not do, for example regulating prematurely and ignoring the positives for the sake of exaggeration. He advises everyone interested in the field to make a change through their professional careers or by donating money. He finishes the last part of his book by drawing upon the imagination of a humanity colonising the universe and maybe even changing its nature to reach the next stage in evolution, if needed. The humans of tomorrow need a chance to fulfil all the things we today can only dream of.

Ord presents an exciting and very good introduction for all those interested in the field of existential risks. He writes eloquently and yet very understandable, avoiding technical terminology wherever possible while explaining it well whenever he can't, making it an altogether interesting read even for a non-academic audience. The structure of the book is inherently sound and his overall tone of voice sounds calm and rational. And yet, this very interesting book is not without its flaws.

First, Ord leaves out a major part of philosophical debates revolving around population ethics, dedicating only a few pages in the appendices to it. The book could have benefited immensely from this if it dived deeper into the debates of human nature, ethics, population and potential. Especially the debate around s-risks (risks of astronomical suffering) that explain how a future does not only have to include happiness but also an huge amount of potential suffering could have been helpful. S-risks put into question whether extinction would be the worst scenario if the alternative would be to cause unprecedented amounts of suffering. Thus in some scenarios, we could not find ourselves in an existential risk, but a s-risk. Lowering the existential risk could therefore raise the s-risks. How then do we avoid existential and suffering risks and still find the best future? Ord argues that we constantly made progress, fighting poverty, strengthening women's rights, and making education possible for more humans than ever before, but that this does not guarantee our steady progress in the future. We could still evolve back on issues or never find a consensus on important subjects. The Long Reflection part of the book is made out to be the time when humanity finally gets its act together and decides its path in unison – but we should already be talking about all these important issues now, because they determine the way we will walk. Hence we should not worry about bringing people into existence first, but worry about whether those people can live a life worth living. Ord could have given his opinion on the procreation asymmetry and how this influences longtermism and dealing with existential risks.

Second, Ord spends a lot of time on the danger of unaligned AI for humanity, but he neglects the dangers of such a powerful existence for the universe. An unaligned AI could theoretically not only control our planet but decide to colonize space and extend its influence into the galaxy causing irreparable damage and suffering not only for us, but also for other sentient beings, if they exist.

Third, Ord talks about the potential of humanity as if it were an individual, but it is a collective. There is not „one humanity" with its intentions and hopes, but instead people hold many different views and values. He imagines the potential of humanity to be one of high art and science, but one inevitably wonders about the negative potential mankind has also shown to possess, its aggressiveness. Every year we kill billions of animals as a food resource, we wage war against each other and still allow people to starve to death in some parts of the world. What potential for inflicting pain might we possess in the future?

Fourth, Ords view on longtermism, deciding what to do depending on the long-term effects, may be logical from the viewpoint of existential risks, but it could come with catastrophic consequences for present people. For example, if you had to let millions of people suffer now so that in the long-term humanity could benefit from it, you would be inclined to let it happen. But are we not morally obliged to stop suffering whenever we encounter it? Does the suffering of now really pale compared to the happiness of tomorrow? And what kind of quality does the happiness of the future hold, if it was at least partially founded on the sorrow of the past? Talking about the trillions of potential humans in the future suggests that a few million who suffer now don't matter as much, but they do. They are real, they exist and they suffer in contrast to the non-existent humans of the future. There is a real danger of trivialising human lives for the sake of the big picture. Climate change will likely not be the end of humanity, but it will still bring immeasurable pain and suffering to many people, if not stopped – but still this does not make it an existential risk for Ord. But I argue it is an existential risk for all those who will die because of it, will lose land and family and lose hope for the future because of it.

Fifth and finally, the chapter on the risk landscapes seems at times a bit problematic. Ord believes, all things considered, that our odds of facing an existential risk in the next century stand at 1 in 6. Yet, we are talking about risks that have never occurred and that can often only be estimated in rough ways, or that could potentially be much bigger or lower than we might dare think. Ord admits that all of his estimates are just his best guesses and should not be taken as precise mathematics, but those evolutions need a stronger ground on which to base our actions on if we were to take existential risks more seriously. We will need more work on the field of risk theory to better understand existential risks.

In the end, Toby Ord has delivered a very compelling book on one of the most interesting and maybe underrepresented subjects in the public discourse. He manages to give a well written introduction into existential risks, even though it ignores a large spectrum of philosophical debate, but leaves the reader wanting to learn more about our potential and the risks we could face. Its maybe biggest accomplishment is to give the reader a sense of hope, even in the face of our potential doom. One can only agree with Ord, that things are always largely in our hands.

Ord, Toby (2020): The Precipice. Existential Risk and the Future of Humanity. London: Bloomsbury Publishing. 480 pages. ISBN 9781526600219 (hardback), ISBN 9781526600233 (paperback), ISBN 9781526600196 (e-book). Price: hardback $34.00/£25.00; paperback $14.95/£10.99; e-book $11.96/£8.79