At the Precipice Now, in Eternal Safety Thereafter?

Review of Toby Ord, *The Precipice: Existential Risk and the Future of Humanity*, Bloomsbury, London (softcover) and Hachette, New York (illustrated), 2020, 480 pp., ~30 USD.

Simon Friederich and Emilie Aebischer

University of Groningen

University College Groningen

Hoendiepskade 23/24

9718BG Groningen

The Netherlands

s.m.friederich@rug.nl and e.aebischer@student.rug.nl

In *The Precipice*, Toby Ord takes a broad sweep through the risks threatening human civilization to argue that we have arrived at a special point in history. According to him, civilization is standing at the Precipice, a time of unprecedented levels of risks of extinction and unrecoverable collapse, but also a time of unique opportunity to navigate the dangerous path that leads to the other side, which holds great rewards.

Ord's thinking is shaped by "effective altruism" -- an approach to doing good that is pursued by a vibrant community of individuals who try to invest the resources at their disposal in the most effective way to achieve the most beneficial outcomes possible. A key priority identified by effective altruists is improving humanity's longterm perspective. Ord sides with those who see many promising trends in human development such as declines in extreme poverty and child mortality and increases in literacy and life expectancy. If these trends continue and adverse trends such as environmental degradation are reversed, civilization may be on track for unprecedented flourishing, potentially on a scale of billions of years.

Ord refers to the historical possibility of such scenarios of unprecedented flourishing as "humanity's longterm potential", and characterizes an existential risk as "a risk that threatens the destruction of humanity's longterm potential" (p.37). An existential catastrophe would thus be an existential risk come true, either in the form of extinction or by locking us into, as Ord puts it, "an unrecoverable dystopia – a world with civilization intact, but locked into a terrible form, with little or no value" (p.153). According to Ord, there are very serious risks threatening to destroy humanity's longterm potential today and in the coming centuries. Combining this with the view that humanity's future potentially is of immense value, he and many other effective altruists conclude that combating existential risks should be among our highest priorities.

Ord divides existential risks into three categories: natural risks, current anthropogenic risks and future anthropogenic risks. Natural risks can be estimated based on the historical record of asteroid and comet impacts, supervolcanic eruptions, and stellar explosions. Ord assigns a credence of one in 10'000 chance to human extinction from any of these risks in the next

hundred years. He makes a similar estimate for the risk from natural pandemics, which is differentially increased and decreased by recent historical developments such as enhanced international mobility on the one hand and progress in medicine and biotechnology on the other.

Technological progress is behind many promising historical trends and it may allow us to virtually eliminate natural risks. But, paradoxically, as it continues and helps us leave behind poverty and reduce natural risks, it creates new -- and, according to Ord, larger -- existential risks. Indeed, Ord -- very pessimistically here, as we see it -- assigns a one in six chance to an anthropogenic risk leading to an existential catastrophe in the next century. The five main current and future anthropogenic risks that he identifies are nuclear war, climate change, environmental degradation, engineered pandemics and "unaligned" artificial intelligence, i.e. advanced artificial intelligence that either does not behave as predicted or behaves as predicted, but with devastating consequences that its creators failed to take into account.

Some of these risks, for example climate change or environmental degradation, may function mostly as "risk factors" by making other risks more severe. A further risk factor is conflict between major military powers, which increases the risks from war with nuclear weapons and bioweapons.

The book culminates in what Ord calls a "grand strategy for humanity". This is a three step plan which consists of (1) safeguarding humanity's potential by decreasing existential risks to negligible levels ("achieving existential security"), (2) "the Long Reflection", a time during which human civilization reflects on and decides the kind of future it wants (e.g. whether it

wants to modify our biological makeup or branch into distinct sub-civilizations), and (3) realizing our potential.

What we can hope to achieve in our century, according to Ord, is to move forward with the first step: to navigate the Precipice and steer humanity into relative safety. The significance of this task is enormous according to Ord, who estimates that "about a third of the existential risk over our entire future lies in this century." (p.170) Strikingly, his view combines severe pessimism about our current situation with bright optimism about civilization's prospects in the longer term. Stunningly, he thus argues that, if civilization makes it through the next few centuries, it will likely have reached existential security and persist indefinitely by standard historical time scales.

This view is remarkable especially because of the special position it assigns to humans today in the grand scheme of history. When combined with the belief that we can have a significant effect on history, it entails the striking claim that we are amongst the most influential agents ever, living at the very point in time where it gets decided whether our civilization exists for a mere moment of cosmic time or for eons. As MacAskill (forthcoming) persuasively argues, psychological traps such as salience -- our own actions seem particularly important to us -- and confirmation bias may cause us to believe such a claim even if we do not have any strong evidence for it.

And indeed there are good reasons to doubt that we are amongst the most influential agents ever. For example, both the currently uncomfortably high level of existential risk and our ability to reduce it stem in large part from technological progress. It seems natural to expect both that

technological progress will confront future humans with novel risks and that those future humans will have even more powerful tools to combat them. Unfortunately, we see no principled reason for confidence that, as a net outcome, risks will decrease drastically. Moreover, existential security may not be permanent. While we might have reached *a* precipice and may manage to overcome it, future civilizations could face novel, potentially more dangerous, ones.

There is one reason for optimism when it comes to humanity's capability to decrease existential risks: namely, we are only now starting to systematically tackle them. One may hope that they will considerably decrease once benevolent, rational, and well informed actors realize what is at stake and act accordingly.

However, we worry that even such actors may not be able to reduce existential risks to negligible levels. Our worry is fuelled by the observation that the most significant existential risks stem in large part from collective action problems that make unilaterally driven mitigation hard. For example, the threat of nuclear war arises from prisoner's dilemma-style arms race between military powers. The threat of climate catastrophe also arises from such a problem: it is in international actors' individual interest to derive relatively cheap and flexible energy from fossil fuels, and this creates the "tragedy of the commons" problem of an atmosphere with increased CO2 concentration, resulting in climate change.

An open question is whether *any* such risk can be eliminated if some actors take intelligent action even if others refuse to cooperate. For the climate problem, part of the most promising approach is to advance the deployment of low-carbon energy sources and make them more

competitive. This may ultimately cause such sources to outcompete fossil-fuel based ones, resolving the collective action problem. Ord thinks greatly of the potential of solar power and ultimately expects its large-scale deployment in interplanetary space, harvesting energy from the sun that would otherwise be "wasted" (p. 228). Here again we worry that Ord's unbridled optimism -- contrasting with his pessimism about current risks -- could turn out premature. History is rife with unfulfilled expectations about specific technologies, and whether solar power in space will be energetically advantageous and economically profitable is, in our view, not knowable today.

Ord provides useful concrete policy suggestions concerning the risks considered (Appendix F). His general recommendation for dealing with potentially risky technologies is that we should "speed[...] up the development of protective technologies relative to dangerous ones" (p. 208). What this means in practice, however, is not always obvious. To illustrate this, we briefly discuss an idea that Ord toys with to reduce the risks from nuclear war: a temporary ban on "nuclear technologies until we've had a hundred years without a major war" (p. 207).

A first difficulty with this idea is what to count as a "nuclear technology." Notably, nuclear medicine is beneficial to humans and related to nuclear weapons only in the sense that the underlying physical principles overlap. Banning it would probably not be effective in reducing the risk of nuclear war and do significant harm. Banning civilian nuclear energy might make more sense because its infrastructure, particularly uranium enrichment facilities, could be diverted to help produce nuclear weapons. But such a ban should not be implemented lightly: besides solar energy, nuclear energy is one of the few highly scalable low-carbon sources with

large climate mitigation potential. It was crucially involved in some of the fastest historical electricity decarbonization episodes, notably, in France, Sweden, and Ontario: banning it may lower our chances to achieve decarbonization in this century and increase climate risks. Moreover, if imposed and implemented by only a limited number of actors, the ban may do little to mitigate weapons proliferation while substantially diminishing climate mitigation potential. As suggested by Gibbons (2020), striving for technological leadership in civilian nuclear technology and contributing to the worldwide dissemination of proliferation-resistant practices might be more effective overall. We expect that nearly all very powerful and beneficial technologies raise similarly intricate trade-offs when deployed at scale.

Further factors may complicate Ord' conclusions and recommendations. As he insists, to safeguard its potential, humanity should keep its options open: lock-in events may prevent it from fulfilling its potential and are, in that sense, existential risks themselves. But, as Ord acknowledges (p. 387), this maxim is in tension with the imperative of existential risk minimization, as stabilizing some risks may require us to restrict future possibilities. This tension could take uncomfortable forms: as Bostrom (2019) argues, we may develop "black ball" (p.455) technologies which will instantaneously and drastically increase the likelihood of extinction. The stabilization of these risks might require "ubiquitous-surveillance-powered preventive policing" and "effective global governance" (p.467), which in turn would open up avenues for governments to control and suppress individuals. Indeed, the idea of reaching a state of permanent existential security can be seen as equivalent to "locking in" a historical trajectory without existential risks. In the best case, this merely complicates Ord's narrative. In the worst case, it may mean that reaching a state of affairs which we might consider dystopian -- and

which he might rank as an existential catastrophe -- could one day offer our only realistic chance of avoiding extinction.

Our critical reflections notwithstanding, we wholeheartedly welcome Ord's book as an urgently needed stimulus for the academic community to tackle existential risks systematically. One of its biggest achievements is that it brings into focus how frighteningly little knowledge we have when it comes to existential risks and how to respond to them. By offering a compelling panorama of what is at stake Ord sets the stage for future attempts to enhance and deepen our understanding of those risks and, ultimately, act accordingly.

References

Bostrom, Nick. 2019. The vulnerable world hypothesis. Global policy 10:455-476.

Gibbons, Rebecca Davis. 2020 Supply to deny: the benefits of nuclear assistance for nuclear nonproliferation. Journal of Global Security Studies, 5:282-298.

MacAskill, William. forthcoming. Are We Living at the Hinge of History?. In Ethics and Existence: The Legacy of Derek Parfit, ed. Jeff McMahan, Tim Campbell, James Goodrich, and Ketan Ramakrishnan. Oxford: Oxford University Press.