The Precipice – Existential Risk and the Future of Humanity. Ord, T., 2020 London, Bloomsbury Publishing. 480 pp, £22.50 (hd)

Sand, M.

DOI
10.1111/japp.12512

Publication date
2021

Document Version
Accepted author manuscript

Published in
Journal of Applied Philosophy

Citation (APA)
Sand, M. (2021). The Precipice – Existential Risk and the Future of Humanity. Ord, T., 2020 London, Bloomsbury Publishing. 480 pp, £22.50 (hd). *Journal of Applied Philosophy*, *38*(4), 722-724. Article 10.1111/japp.12512. https://doi.org/10.1111/japp.12512

This is a pre-print version of a review that has been published in the *Journal of Applied Philosophy* (**DOI: 10.1111/japp.12512**) – please cite the printed version.

The Precipice - Existential Risk and the Future of Humanity
T. Ord, 2020
London: Bloomsbury Publishing
480 pp, £22.50 (hd)

***

Martin Sand

***

In *The Precipice*, Toby Ord conveys a serious message. Mankind stands at a figurative cliff's edge, a crucial moment in the history of our species: "Humanity lacks the maturity, coordination and foresight necessary to avoid making mistakes from which we could never recover. As the gap between our power and our wisdom grows, our future is subject to an ever-increasing level of risk. This situation is unsustainable." (pp. 3-4) Ord suggests that human extinction is a distinctively bad event, in some sense, worse than the badness of hundreds of millions of people dying. Human extinction according to Ord also encompasses states in which humanity falls short of reaching its full potential (p. 41), for instance, by getting locked-in in an "unrecoverable" dystopian state. The main reason for the badness of human extinction and such states is that humanity would be deprived of "almost all our potential for a worthy future." (p. 154) In

contrast to previous writers discussing the ethics of climate change and the prospect of nuclear fallouts Ord introduces a novel ground for re-thinking responsibility for the future: The unrecoverable collapse of civilization would not only entail enormous misery, but deprive humanity of its potential and this is a particular evil that has not been take seriously thus far.

The Precipice is the state in which humanity is "at high risk of destroying itself." (p. 33) Existential risks are not only human-borne: They also include asteroid impacts, supervolcanic eruptions and stellar explosions. Ord considers those natural risks to be rather low. More important for our survival are anthropogenic risks, a list that is led by the existential threat through nuclear weapons. Ord suggests that a "nuclear winter appears unlikely to lead to our extinction." (p. 99) As elsewhere, he grants that there are significant uncertainties regarding our understanding such scenarios (p. 100). Anthropogenic risks also encompass climate change and pandemics, whose likelihood have increased due to extensive livestock farming and urbanization (p. 126) – an insight that receives much attention since COVID-19. Climate change, too, is listed in *The Precipice* as a man-made phenomenon, whose understanding is infused with uncertainty. Lastly, there is the threat of Artificial Intelligence (AI), which receives quite some attention in *The Precipice*.

AI systems that stick at nothing to maximize reward for achieving some predetermined (or self-determined) goal and, thereby, seizing control over humanity is according to Ord "the most plausible existential risk from AI." (p. 148) Ord believes that there are many ways for an AI system to escalate its power and seize control. The technology doesn't have to emerge in the form of robots to be existentially threatening: "So long as an AI system can entice or coerce people to do its physical bidding, it wouldn't need robots at all." (p. 146) Ord does not hesitate to consider this as "the most speculative case for a major risk in [*The Precipice*]." (p. 149) Why then does he belief in the plausibility of this scenario and goes so far as to estimate the chance that unaligned AI will lead to human extinction within the next 100 years as roughly 1 in 10 (p. 167)? Ord suggests that it is useful to listen to AI experts and he elicits that "many AI researchers take seriously the possibilities that AGI [artificial general intelligence] will be developed within 50 years and that it could be an existential catastrophe." (p. 151) Such expert reliance is clearly problematic, as they anticipate the future of their own respective fields with little understanding of the developments of related fields (economics, politics, society etc.) that can have an enormous impact on actual developments. The success or failure of AI rests, amongst others, on peoples' (politicians, citizens and stakeholders) desires and concerns regarding this technology and AI experts might hugely misinterpret

those. For those and others reasons, expert-based predictions (e.g. Delphi-methods) have often been criticized. Despite the repeated admittance that much guess-work is involved in risk assessments, Ord does not shy away from putting a precise probability to the total existential risk (~1 in 6 chance within the next 100 years). It is a stark omission that the much-contested concept "risk" remains underdeveloped throughout the book and seems all too often identified with mere probability.

Ord's argument for the evil of extinction is one about deprivation: Like a person who prematurely dies and is, thus, deprived of her future, humanity would be deprived of reaching its potential, if civilization collapses. But, humanity – unlike persons – doesn't have wishes for the future and the concept "potential" is value neutral. This gives rise to typical challenge for deprivation arguments: What is good about survival, if undesired and undergone in a state of agony? Looking back on the short history of humanity, Ord detects social, technological and moral progress, an increase of wealth and longevity and he, thus, seems to extrapolate those positive developments, subliminally foisting a positive connotation into the concept of "humanity's potential" that is not a genuine part of its semantics. If the deprivation argument is not about potentiality *per se* (Ord discusses the potential for totalitarianism and considers this as an undesirable outlook), one might ask, precisely which potential is so

valuable to be sustained at almost all cost? What is Ord's positive vision of humanity's potential? As a response, Ord succumbs to technological escapism not unfamiliar from the technophile literature: Humans might travel to other stars and save large parts of the biosphere, bringing seeds and cells with them (p. 223): "we could harness the [sun's] energy by constructing solar collectors […]." (p. 228) This is certainly an interesting prospect, but why does it matter? The question about the notion and value of humanity's potential is important: While Ord considers creating "Existential Security" as important as the "Long Reflection" – thinking about the governance of our societies and about the future of human values – he thinks the former to be more urgent in *The Precipice*. Such ranking seems frail, if mere survival doesn't always trump quality. As is the case with regard to persons, humanity's survival is interwoven with questions about the quality of humans' lives and deserve to be considered simultaneously.

At the beginning, Ord suggests that there hasn't been enough attention to the issue of extinction since the end of the Cold War and that many people, therefore, underestimate those risks (p. 42). On good grounds, Eva Horn argues for the opposite in her *Future as Catastrophe - Imagining Disaster in the Modern Age* (New York: Columbia University Press, 2018): We are living in heydays of thinking about "the end of humanity". This is verified by a bourgeoning debate about speculative technological futures, a source of literature that has been

woefully neglected in *The Precipice*. In Futures Studies, Science and Technology Studies and philosophy of technology, many reasonable suggestions to deal with technological risks have been brought forward in the past years including Constructive Technology Assessment, and Responsible Research and Innovation. Those are concrete frameworks to align technological development with societal values and to support the inclusive envisioning of worthwhile human futures. Ord would have certainly written a different book on existential risk, had he delved into these fields before writing *The Precipice*. This criticism notwithstanding, Ord has written an astonishingly readable and insightful book on a subject that is dizzying multifaceted.

Martin Sand
Delft University of Technology