Book Review of: Toby Ord, *The Precipice: Existential Risk and the Future of Humanity*, Hachette Books, New York, 2020.

Reviewed by: Seth D. Baum, Global Catastrophic Risk Institute
http://sethbaum.com * http://gcri.org

Published in *Risk Analysis*, vol. 42, no. 9, September 2022, pages 2122-2124, DOI 10.1111/risa.13954
This version: 28 October 2022

With COVID-19 raging and climate change intensifying, catastrophic risk is a timely topic. Ord's book should find a receptive audience, but it is not an of-the-moment work. Instead, it fits in the tradition of scholarship that interprets catastrophic risk in terms of the *very* big picture of humanity's role in the universe. Previous books on this theme include Leslie (1996), Rees (2003), Posner (2004), Häggström (2016), and Walsh (2019). Compared to these works, *The Precipice* covers some familiar ground, especially its detailed descriptions of specific risks and its astronomically big picture perspective. It also breaks some new ground, especially its discussion of ethics and its use of quantitative risk analysis.

Of the many books on extreme catastrophic risk, *The Precipice* compares favorably. It tells the story of extreme catastrophic risk as catastrophes so big, they ruin the entire, vast future of civilization. This story traces to work by Ng (1991), Tonn (1999), and Bostrom (2002), and has been explored in a variety of more recent scholarship. Other books touch on this story, but none cover it as carefully as *The Precipice*. *The Precipice* further stands out for its depth of discussion, its quality of scholarship, and its readability. Ord self-describes as a philosopher, but the book is much more than that, and indeed it serves a testament to the importance of interdisciplinary perspectives in risk analysis.

The book is likewise commendable as a work that serves well as an introduction to the topic for a general readership and as a work that will challenge the thinking of experts. The book could be used for graduate or even advanced undergraduate courses as long as students are given some guidance on skimming or skipping the more difficult portions. Exactly which portions would be difficult will vary from classroom to classroom or from reader to read given the book's interdisciplinary nature. For example, some readers may struggle with the book's quantitative risk analysis, whereas others may struggle with its moral philosophy. Readers of all backgrounds should be able to gain some appreciation of the book's perspective on extreme catastrophic risk, which is its primary contribution.

While the book has significant merit, it does get itself into some trouble, especially in its quantitative risk analysis. The book is also unfortunately thin in its discussion of risk management solutions. These are significant concerns, as elaborated below. However, these concerns should be interpreted in light of the overall quality of the book. Before detailing these concerns, here is an outline of the book.

After opening remarks, the book makes its case for the importance of extreme catastrophic risk. The basic idea is that an extreme catastrophe could destroy the entire future of civilization on Earth and potentially into the cosmos. This idea is usually grounded in expected utility consequentialism, but the book also explores other ethics perspectives, which is a welcome contribution. Other perspectives include one generation's duties to others, the virtues of supporting humanity as a whole, and responsibilities to the entire universe. The book then defines the precipice as the "time where humanity is at high risk of destroying itself" (p.40).

Subsequent chapters, constituting about half the book, analyze the risks. Discussions of specific risks are divided into "natural risks", with emphasis on asteroids, comets, supervolcanoes, supernovae,

and gamma ray bursts, “anthropogenic risks”, with emphasis on nuclear weapons, climate change, and other environmental change, and “future risks”, with emphasis on pandemics (both natural and engineered), artificial intelligence, and various scenarios in which humanity locks in some sort of dystopia. For each risk, there is extended description of the risks and some quantitative risk analysis. For most risks, there is also brief discussion of risk management options. The risk analysis culminates in a chapter quantifying the overall space of risks. This contains the author’s subjective probability estimates, which are contentious as discussed below. It also contains nice discussion of other topics, such as how to account for the fact that humanity cannot go extinct twice.

Next is a chapter on risk management options. It begins with broad brush strokes about the overall long-term strategy humanity should take and then goes into somewhat more detail on matters such as the need for anticipatory governance and the role of international cooperation, though the discussion is overall lacking in specificity. Finally, the book closes with another, more detailed discussion of the large value civilization could accrue by persisting into the distant future and expanding beyond Earth.

The organizing concept of *The Precipice* is existential risk. Following Ord’s Oxford colleague Bostrom (2002), *The Precipice* defines existential risk as risk of events that would destroy humanity’s long-term potential. This includes human extinction and scenarios in which humans survive but fail to accomplish the astronomically massive upside of a young civilization in a vast universe. Two concerns with this definition are apparent. The first is semantic. “Existential risk” implies risk to the existence of something, which makes sense for human extinction but not for survival-with-failure. In the latter scenario, humanity still exists; it just isn’t accomplishing as much. I have preferred the term “global catastrophic risk” in part for this reason.

The second concern is analytic. The book’s definition is rooted in a binary: either humanity accomplishes its potential or it doesn’t. Here lies a deeper issue with the book’s analysis that is worth unpacking. The book is certainly correct to emphasize massive scale of humanity’s potential. It is indeed a small world after all, but it is a very, very large universe. If one takes the moral position of caring equally about outcomes regardless of when and where they occur (Ord and I are among the many who do), then this astronomic upside weighs rather heavily in the decision calculus.

Some problems arise when compressing this upside into a yes/no binary. First, not all astronomical upsides are of equal value. *The Precipice* acknowledges this but then punts on the issue, instead calling for a “long reflection”, a period of time to occur after existential risk has been reduced to some minimal level. During this period, humanity is to think through and reach some sort of working consensus on what to do with the universe. This reflection period is proposed on grounds that the risks are more urgent; perhaps they are, but the matter is unsettled. It is additionally unclear how well a long reflection would work in practice or if such a regime could or would even be established in the first place. Given the stakes, it may be prudent to develop contingency plans.

The Precipice additionally applies a strict binary to the severity of catastrophes. Unless an event causes human extinction or fits into a narrow set of other permanent harm scenarios, it is deemed unimportant per the book’s framework. Notably, the collapse of global civilization is rated as unimportant on grounds that civilization is very likely to make a full recovery. The fate of post-collapse populations is a crucial parameter for the analysis of extreme catastrophic risks. The book’s position that recovery is very likely is perhaps the dominant factor in its risk analysis, yet it is given just a single paragraph (p.47) and some brief mentions in footnotes. Other, more detailed analysis has reached less rosy conclusions on the fate of survivors (Baum et al. 2019). This is a complex and deeply uncertain topic in which very little work has been done. The book would have been wise to approach it with greater care and humility.

The book’s emphasis on extreme scenarios is compounded by its strong belief in the resilience of civilization to catastrophes. This is apparent, for example, in its discussion of climate change. The book

considers a scenario in which high degrees of warming result in large portions of Earth becoming physiologically uninhabitable for humans by creating temperature and humidity conditions that exceed the thermal limits of mammals (Sherwood and Huber 2010). However, the book posits that, even at 20ºC of warming, "there would remain large areas in which humanity and civilization could continue", making this not an existential risk (p.114). That is a rather bold claim and is utterly out of step with the current international climate policy debate, which centers on whether to aim for a limit of 1.5ºC or 2ºC. In fairness, the book does otherwise take climate change seriously. It calls for investment in risk management via greenhouse gas emissions reduction and geoengineering, including evaluation of the risks geoengineering poses. It additionally calls for research to clarify the extreme risks that climate change may pose (see also Beard et al. 2021). Nonetheless, it seems massively optimistic to dismiss 20ºC as no big deal.

The same "extremist" perspective underlies the rest of the book's risk analysis. It places the most weight on runaway artificial intelligence (AI) scenarios in which AI takes over the world and kills everyone. In these scenarios, resistance is futile and humans rapidly go extinct, leaving no chance of civilization recovery. There's nothing wrong with believing in the importance of AI risk; I for one agree that the risk is significant. But its comparative importance in the book appears to be an artifact of the book's strong belief in the unimportance of civilization collapse scenarios, which factor centrally in most other extreme catastrophic risks. The book does acknowledge that "humanity may be more vulnerable following a global catastrophe" (p.176), but this point is not well integrated into its overall analysis.

The book's comparative risk analysis raises methodological questions that are likely to be familiar to a risk analysis readership. In particular, the book includes a compilation of estimates for the probability of existential catastrophe for each of the risks it covers (Table 6.1). The estimates are described as Ord's personal subjective probability estimates in consideration of the evidence presented in the book and Ord's broader knowledge of the topic. These estimates are prefaced by some words of caution on the imprecision of subjective estimates of deeply uncertain parameters and justified on grounds that numbers are needed "to reason clearly about the comparative sizes of different risks, or classes of risks" (p.164). The numbers are then used to inform the book's various recommendations.

To quantify or not to quantify is a quintessential risk analysis question. On one hand, risk estimates can inform decision-making, and there is a perspective in which having some number, no matter how tentative, is better than having none. On the other hand, poorly formed estimates can induce bad decision-making via overconfidence and other biases. The subjective judgments of a single individual, no matter how well-informed, are arguably among the least rigorous ways to quantify uncertainty (Beard et al. 2020), especially for complex, unprecedented topics (such as extreme catastrophic risk) in which there may be no one who classifies as an expert (Morgan 2014). Furthermore, it is common for risk estimates to be used uncritically without regard for the biases and uncertainties that may underlie the estimates. I have already seen uncritical use of the estimates in *The Precipice* on multiple occasions, including by people whom one may expect to know better. I am very sympathetic to Ord's desire to inform sound decision-making through the inclusion of subjective risk estimates, and I appreciate the considerable care that was clearly taken in producing the estimates and describing their limitations. I likewise understand the sense in which omitting the estimates leaves readers in the dark. However, for risks as deeply uncertain as these, perhaps the darkness is the point.

The pitfalls of quantification is one area in which greater attention to risk management solutions would have been helpful. For guidance on how to address the risks, quantification is not always needed. For example, the book's recommendation of reducing greenhouse gas emissions does not depend on the severity of (for example) 2ºC or 20ºC of warming: regardless of the details, emissions reduction is still going to be a prudent course of action, at least for a much larger amount of emissions

reduction than is currently being pursued. The quantitative comparative risk analysis emphasized in the book is mainly helpful for decisions involving multiple risks, in particular how to allocate scarce resources across the risks and how to make risk-risk tradeoffs (e.g., Baum 2019). However, even these decisions can be well informed by analysis that stops short of producing "the answer" to quantitative questions. Furthermore, these cross-risk decisions are not covered in the book. Instead, the book's extensive and contentious quantification was, at least for its own purposes, unnecessary.

The book's discussion of risk management solutions would have further benefited from consideration of the perspectives of the relevant actors. There is an acute tension between the book's unusually extreme perspective on catastrophic risk and the more moderate perspectives that are more commonly held. The book laments the lack of attention to extreme catastrophic risk, but its primary solution is to present arguments from moral philosophy. Such arguments can sometimes be persuasive, but often they are not. Thus, there is a need to develop solutions that reduce extreme catastrophic risk and that can appeal to actors for other reasons (Baum 2015). Greenhouse gas emissions reduction is one example: emissions reduction also reduces risks from more moderate forms of climate change and often involves significant co-benefits such as improved local air quality. As a general matter, social and political viability should be primary considerations for the development of solutions for managing extreme catastrophic risk, but they do not factor into the book's analysis.

These sins of omission and analytical shortcomings imply that *The Precipice* is not the final word on extreme catastrophic risk. Nonetheless, the book is an excellent contribution to a very important topic.

Acknowledgments
Tony Barrett and Editor Michael Greenberg provided helpful feedback on an earlier draft. Any remaining errors are the author's alone.

References

Baum, S. D. (2015). The far future argument for confronting catastrophic threats to humanity: Practical significance and alternatives. *Futures*, *72*, 86-96.

Baum, S. D. (2019). Risk-risk tradeoff analysis of nuclear explosives for asteroid deflection. *Risk Analysis*, *39*(11), 2427-2442.

Baum, S. D., Armstrong, S., Ekenstedt, T., Häggström, O., Hanson, R., Kuhlemann, K., et al. (2019). Long-term trajectories of human civilization. *Foresight*, *21*(1), 53-83.

Beard, S., Rowe, T., & Fox, J. (2020). An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards. *Futures*, *115*, 102469.

Beard, S. J., Holt, L., Tzachor, A., Kemp, L., Avin, S., Torres, P., & Belfield, H. (2021). Assessing climate change's contribution to global catastrophic risk. *Futures*, *127*, 102673.

Bostrom, N. (2002). Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Evolution and Technology*, *9*(1).

Häggström, O. (2016). *Here Be Dragons: Science, Technology, and the Future of Humanity*. Oxford: Oxford University Press.

Leslie, J. (1996). *The End of the World: The Science and Ethics of Human Extinction*. London: Routledge.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National academy of Sciences*, *111*(20), 7176-7184.

Ng, Y. K. (1991). Should we be very cautious or extremely cautious on measures that may involve our destruction? *Social Choice and Welfare*, *8*(1), 79-88.

Posner, R. (2004). *Catastrophe: Risk and Response*. Oxford: Oxford University Press.

Rees, M. (2003). *Our Final Century: Will the Human Race Survive the Twenty-First Century?* Oxford: William Heinemann.
Sherwood, S. C., & Huber, M. (2010). An adaptability limit to climate change due to heat stress. *Proceedings of the National Academy of Sciences*, *107*(21), 9552-9555.
Tonn, B. E. (1999). Transcending oblivion. *Futures*, *31*, 351-359.
Walsh, B. (2019). *End Times: A Brief Guide to the End of the World*. New York: Hachette Books.