briefing note 415 May 2020

# Ranking of Potential Causes of Human Extinction

Paul Somerville, Risk Frontiers

We are good at learning from recent experience; the *availability heuristic* is the tendency to estimate the likelihood of an event based on our ability to recall examples. However, we are much less skilled at anticipating potential catastrophes that have no precedent in living memory. Even when experts estimate a significant probability for an unprecedented event, we have great difficulty believing it until we see it. This was the problem with COVID19: many informed scientists (e.g. Gates, 2015) predicted that a global pandemic was almost certain to break out at some point in the near future, but very few governments did anything about it.

We are all familiar with the annual *Global Risk Reports* published by the World Economic Forum. Looking at their ranking of the likelihood and severity of risks (see Figure I, page 1 of the 2020 report), we see that the rankings over the past three years have consistently attributed the highest likelihood to *Extreme Weather* events and the highest impact to *Weapons of Mass Destruction*. However, in 2020, *Climate Action Failure* displaced *Weapons of Mass Destruction* as the top impact risk. Further, the rankings have changed markedly over the past 22 years, and while it may be that human activity has had an inordinately large impact on objective risks levels such as that due to *Weapons of Mass Destruction* in the last three years, there is probably a large component of subjectivity and *availability heuristic* in the rankings reflecting changing risk perceptions.

The work of Toby Ord and colleagues described below stands in stark contrast with these risk assessments. First, it addresses much more dire events that could lead to human extinction. Second, it attempts to use objective methods to assess the risks to avoid problems arising from risk perception. This work results in some surprising and thought-provoking conclusions, including that most human extinction risk comes from anthropogenic sources other than nuclear war or climate change.

Australian-born Toby Ord is a moral philosopher at the Future of Humanity Institute at Oxford University who has advised organisations such as the World Health Organisation, the World Bank and the World Economic Forum. In *The Precipice*, he addresses the fundamental threats to humanity. He begins by stating that we live at a critical time for humanity's future and concludes that in the last century we faced a one-in-a-hundred risk of human extinction, but that we now face a one-in-six risk this century.

In previous work, Snyder-Beattie et al. (2019) estimated an upper bound for the background rate of human extinction due to natural causes. Beckstead et al. (2014) addressed unprecedented technological risks of extreme catastrophes, including synthetic biology, geoengineering (employed to avert climate change), distributed manufacturing (of weapons), and Artificial General Intelligence (AGI); see also Hawking (2010). In what follows, the conclusions of these studies are summarised and the various potential causes of human extinction ranked (Table 1).

Natural risks, including asteroids and comets, supervolcanic eruptions and stellar explosions are estimated to have relatively low risks, which, taken together, contribute a one-in-a-million chance of extinction per century.

Turning to anthropogenic risks, the most obvious risk to human survival would seem to be that of nuclear war, and we have come near it, mainly by accident, on several occasions. However, Ord doubts that even nuclear winter would lead to total human extinction or the global unrecoverable collapse of civilisation. Similarly, Ord considers that while climate change has the capacity to be a global calamity of unprecedented scale, it similarly would not necessarily lead to human extinction. He also considers that environmental damage does not show a direct mechanism for existential risk. Nevertheless, he concludes that each of these anthropogenic risks has a higher probability than that of all natural risks put together (one-in-a-million per century).

Future risks that Ord considers include pandemics, "unaligned artificial intelligence" (superintelligent AI systems with goals that are not aligned with human ethics), "dystopian scenarios" ("a world with civilisation intact, but locked into a terrible form, with little or no value"), nanotechnology, and extraterrestrial life.

Ord considers the risk represented by pandemics to be mostly anthropogenic, not natural, and the risk from engineered pandemics is estimated to be one-in-30 per century, constituting the second highest ranked risk. He does not consider COVID 19 to be a plausible existential threat.

Ord considers that the highest risk comes from unaligned artificial intelligence. Substantial progress in artificial general intelligence (AGI) could someday result in human extinction or some other unrecoverable global catastrophe. The human species currently dominates other species because the human brain has some distinctive capabilities that other animals lack. However, if AI surpasses humanity in general, then it becomes "superintelligent" and could become powerful and difficult to control. The risk of this is estimated to be one-in-10 per century. These risks combine for a one-in-six chance of extinction per century.

The methodology behind Ord's estimates is described in detail in the book and in the answers to questions he was asked in the 80,000 Hours podcast (2020). For example, for the case of AGI, Ord states that the typical AI expert's view of the chance that we develop smarter than human AGI this century is about 50%. Conditional on that, he states that experts working on trying to make sure that AGI would be aligned with our values estimate there is only an 80% chance of surviving this transition while still retaining control of our destiny. This yields a 10% chance of not surviving in the next hundred years.

In the rankings in Table 1, all considered anthropogenic risks, shown in Roman; exceed all natural risks, shown in italics

**Table 1. Ranking of Risks of Human Extinction**

| Risk | 1/Frequency | Source |
|---|---|---|
| Unaligned AI | 1 kyr | Ord, 2020 |
| Engineered Pandemic | 3 kyr | Ord, 2020 |
| Unforeseen Anthropogenic Risks | 3 kyr | Ord, 2020 |
| Other Anthropogenic Risks | 5 kyr | Ord, 2020 |
| Nuclear War | 100 kyr | Ord, 2020 |
| Climate Change | 100 kyr | Ord, 2020 |
| Other Environmental Damage | 100 kyr | Ord, 2020 |
| *Natural Pandemic* | *1 Myr* | *Ord, 2020* |
| *Supervolcano Eruption* | *1 Myr* | *Ord, 2020* |
| *Flood Basalt* | *32 Myr* | *Snyder-Beattie et al., 2019* |
| *Asteroid or Comet Impact* | *100 Myr* | *Ord, 2020* |
| *Supernova* | *100 Byr* | *Ord, 2020* |

**References**

80,000 Hours (2020). https://80000hours.org/podcast/episodes/toby-ord-the-precipice-existential-risk-future-humanity/#robs-intro-000000

Beckstead, Nick, Nick Bostrom, Neil Bowerman, Owen Cotton-Barratt, William MacAskill, Seán Ó hÉigeartaigh, and Toby Ord (2014). *Unprecedented Technological Risks*. https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf.

Gates, Bill. (2015). The next outbreak? We're not ready. https://www.ted.com/talks/bill_gates_the_next_outbreak_we_re_not_ready/transcript?language=en

Hawking S. (2010), *Abandon Earth or Face Extinction*, Bigthink.com, 6 August 2010.

Snyder-Beattie, Andrew E., Toby Ord and Michael B. Bonsall (2019). An upper bound for the background rate of human extinction. *Nature Reports*, https://doi.org/10.1038/s41598-019-47540-7.

Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury.

Rougier, J., Sparks, R. S. J., Cashman, K. V. & Brown, S. K. The global magnitude–frequency relationship for large explosive volcanic eruptions. *Earth Planet. Sci. Lett.* 482, 621–629 (2018).

World Economic Forum (2020). The Global Risks Report 2020. http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf