Guest editorial

Olle Häggström and Catherine Rhodes

Olle Häggström is based at Chalmers University of Technology, Gothenburg, Sweden. Catherine Rhodes is based at Centre for the Study of Existential Risk, University of Cambridge, Cambridge, UK.

Existential risk to humanity

Existential risks are those that threaten the extinction of humanity or, somewhat more broadly, those "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom, 2002). While it has long been widely recognized that global nuclear war and catastrophic climate change have potentially civilization-threatening consequences, it is only in the last decade or two that scholars have begun, in work that is typically highly interdisciplinary, to systematically investigate a broader range of existential risks. A landmark event was a conference on global catastrophic risks in Oxford in 2008, and the accompanying book edited by Bostrom and Ćirković (2008). Subsequent research has tended to confirm the impression from that event that, in the time frame of a century or so, natural risks (such as asteroid impacts) are out shadowed by anthropogenic ones. In addition, while much work remains to be done in identifying and understanding the various risks, there is at present an increasing focus also on how to avoid or mitigate them.

We are not yet at a stage where the study of existential risk is established as an academic discipline in its own right. Attempts to move in that direction are warranted by the importance of such research (considering the magnitude of what is at stake). One such attempt took place in Gothenburg, Sweden, during the fall of 2017: an international guest researcher program on existential risk at Chalmers University of Technology and the University of Gothenburg, featuring daily seminars and other research activities over the course of two months, with Anders Sandberg serving as scientific leader of the program and Olle Häggström as chief local organizer, and with participants from a broad range of academic disciplines. The nature of this program brought substantial benefits in community building and in building momentum for further work in the field: of which the contributions here are one reflection. The present special issue of Foresight is devoted to research carried out and/or discussed in detail at that program. All in all, the issue collects ten papers that have made it through the peer review process.

As well as this recognition of the necessity of deep interdisciplinarity within their research efforts, the existential risk community has a core focus on how their work can improve humanity's preparation for and capacity to manage and mitigate such risks. Two of the papers outline options for concrete actions; several others point to a range of barriers to achieving informed decisions and effective practical actions in a timely manner. These include problems of recognition or perception of extreme risks (e.g. through cognitive biases) and problems for governance and control that arise both from the rapidity of technological advance and the complexity and vulnerabilities of the human and natural systems in which they are occurring. This collection of papers, therefore, not only contributes to existential risk scholarship, but also to its engagement with decision-making communities.

The issue opens with a survey paper by Phil Torres, who argues that the main existential risks in the coming century can be broadly categorized as:

DOI 10.1108/FS-03-2019-108

- those arising from environmental degradation;
- those that stem from the increasing availability and proliferation of increasingly powerful weapons technologies; and
- those relating to advances in artificial intelligence (AI) and in particular the emergence of so-called machine superintelligence.

Of course there is intersection and interaction between these kinds of risks; Torres emphasizes the need to understand and handle them jointly and offers what he calls the Great Challenges framework for doing so.

Existential risks from environmental degradation also appear in Karin Kuhlemann's paper, which provides a broader view of the context in which existential risks will play out. Many of the existential risk scenarios treated in other contributions to the special issue are extreme in their sometimes science fiction-like character and tend to play out rapidly and dramatically. To the extent that all this drama risks hijacking our attention so that we forget to deal with the more gradual emergencies that are already ongoing, Karin Kuhlemann offers an effective antidote and argues that at the core of these gradually accumulating and (in her preferred term) "unsexy" risks we find the problem of overpopulation.

Another kind of birds-eye view is taken by Seth Baum and his 13 coauthors in their paper on long-term trajectories of human civilization, where "long-term" is meant very seriously as the entire future time span over which humanity may continue to exist, such as millions or billions of years. The possible trajectories may in principle be of various kinds, involving long periods of status quo, catastrophes, radical technological transformations, expansion beyond our current home planet, or combinations of these. The plausibility of these various kinds of trajectories are evaluated, and tentative conclusions are drawn about how events in the near term can affect the long term, as well as about what this means normatively for our current actions and how this depends on the choice of ethical framework.

Other considerations relating to time and timeliness appear in the paper *There's plenty of time at the bottom* whose title is a play with Feynman (1959). Anders Sandberg investigates how, in a world increasingly dominated by artificial computing machinery, drastic speedup of computation may lead not only to unprecedented opportunities for producing great value, but also (to quote from the concluding section of the paper) "to control gaps, systemic risks, speed inequalities, and overly fast or uncertain decisions". While this paper offers a more abstract perspective, it points towards significant challenges for timely decision-making and action on existential risks (particularly those related to emerging technologies).

The next two papers provide some of the preparatory analysis and suggest measures that can be a means of preparing for and responding to such risks in a timely manner, dealing more concretely than other papers on how humanity could go about to survive under various catastrophic conditions. First, Alexey Turchin and Brian Green offer a first systematic treatment of how islands might be used as refuges following a global catastrophe. Second, David Denkenberger, Joshua Pearce, Andrew Ray Taylor and Ryan Black deal with a range of catastrophes that can lead to nuclear winter-type scenarios where sunlight is blocked over a substantial period and lead to global food shortage and famine, and the life-saving potential and cost-effectiveness of various schemes for providing alternate foods to survivors of the catastrophe.

In the contribution by James Miller, a somewhat counterintuitive finding is offered, that two existential risks can be preferable to just one of them, provided the right kind of tension between them. In the example he provides, one of the risks is that of a late filter in the so-called Great Filter framework for understanding the Fermi Paradox (Hanson, 1998; Häggström, 2016), while the other is the risk from creating a nonaligned super intelligent AI.

The latter of these risks brings us to a cluster of three papers dealing specifically with AI-related risk. First, Roman Yampolskiy applies a historic perspective in an attempt to be systematic about what lessons for the future can be learned from previous AI failures. Second, Olle Häggström takes as starting point for his contribution the observation that once a super intelligent AI is in place the future of humanity is likely to hinge on what goals and values are held by the machine, and digs into the challenges involved in trying to predict (and hopefully control) what these values will be. Third, Karim Jebari and Joakim Lundborg offer a distinction between intelligence and the more general notion of *techne*, to argue against the plausibility and likelihood of the kind of very rapid machine takeover associated with terms like *intelligence explosion* and *Singularity*, so that the kinds of apocalyptic AI scenarios discussed in the contributions by Torres, Miller and Häggström may turn out to be less likely than previously suggested.

While we do not claim that the ten papers collected here constitute an exhaustive or even entirely representative look into current existential risk scholarship, we hope that they convey some of the breadth and flavor of the many challenging research problems in this emerging field, and that they will stimulate further discussion and research. We hope also that the positive spirit that pervades it shines through. At the program in Gothenburg, a journalist from Swedish public radio expressed surprise over how happy we seemed to be despite discussing depressing-sounding topics such as the extinction of humanity. The answer to this is that while the problems facing humanity are serious, we are convinced that they are not unsolvable, and we also think about the enormous potential for future human flourishing there will be, provided that humanity gets its act together. To help ensure the latter is what we audaciously aspire to do.

References

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

Bostrom, N. and Ćirković, M. (2008), *Global Catastrophic Risks*, Oxford University Press, Oxford.

Feynman, R.P. (1959), *There's Plenty of Room at the Bottom*, American Physical Society Annual Meeting, Pasadena, CA, available at: www.zyvex.com/nanotech/feynman.html

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

Hanson, R. (1998), "The great filter – are we almost past it?", available at: http://hanson.gmu.edu/greatfilter.html