Working together to face humanity's greatest threats: Introduction to *The Future of Research on Catastrophic and Existential Risk.*

Adrian Currie & Seán Ó hÉigeartaigh

Penultimate Version, forthcoming in Futures

Acknowledgements

We would like to thank the authors of the papers in the special issue, as well as the referees who provided such constructive and useful feedback. We are grateful to the team at the Centre for the Study of Existential Risk who organized the first Cambridge Conference on Catastrophic Risk where many of the papers collected here were originally presented, and whose multi-disciplinary expertise was invaluable for making this special issue a reality. We'd like to thank Emma Bates, Simon Beard and Haydn Belfield for feedback on drafts. Ted Fuller, Futures' Editor-in-Chief also provided invaluable guidance throughout. The Conference, and a number of the publications in this issue, were made possible through the support of a grant from the Templeton World Charity Foundation (TWCF); the conference was also supported by a supplementary grant from the Future of Life Institute. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation or the Future of Life Institute.

1. Introduction

Ours is a resilient species. Around 70,000 years ago our total population may have fallen to between three and ten thousand individuals, possibly due to a supervolcanic eruption (Ambrose 1998)[1]. Yet our ancestors survived, squeezed through the bottleneck, and flourished. But this resilience cannot be taken for granted. We are interconnected and interdependent as never before; the power and scale of our technological capacities are unprecedented. We are in uncharted waters and thus our previous survival is no longer a reason to expect our continued survival (Bostrom 2013). As a result, it is urgent that we develop a systematic understanding of the nature and causes of catastrophic and existential risks.

Human societies are increasingly technologically powerful; it is less certain whether we have developed the foresight and wisdom to steer towards safer futures. For the first time, we might be able to do something about disasters we were previously incapable of predicting, understanding or mitigating: technology and knowledge might free us from some of the natural shocks that human civilisations were once vulnerable to. But these new capabilities themselves generate unprecedented dangers. This calls for the development of research programmes that address the sources of, and possible responses to, such new threats. To do so requires work across academic spaces which is in meaningful dialogue outside of the university. This special issue reflects this aim: it includes new work across disciplines, career-stages and from within and outside academia. Our hope is that it will aid in growing the community of researchers engaged in understanding existential and catastrophic risk, and give a sense of the range of rigorous, fruitful, and urgent research that is being done—and remains to be done—in this area.

Global catastrophic risks have the potential to cause serious harm on a planetary scale (Bostrom and Cirkovic, 2008). A particularly severe category of global catastrophic risk is an existential risk: one that threatens the premature extinction of earth-originating intelligent life, or the permanent and drastic destruction of its potential for future development (Bostrom 2013). The

[1] Although both the timing and extent of the bottleneck, and the role of the Toba volcanic eruption in particular, remain controversial, see Williams (2012).

key difference is that existential catastrophes curtail the possibility of recovery and future development.

Both existential and global catastrophic risks present unique challenges for research and analysis. They represent very rare or even unprecedented developments; therefore there can be little data to draw on, and little opportunity to learn from past events (Currie forthcoming). While some threats are entirely exogenous, many relate to human activity (e.g. climate change, catastrophic biodiversity loss) or the advent of powerful new technologies (e.g. nuclear weapons). The pace of technological progress, and the complexity of the interactions between technology and society more broadly, lead to threats which can be dynamic and difficult to predict (Ó hÉigeartaigh 2017).

Recently, catastrophic and existential risks as a class have become a research target. In 2003 Martin Rees published *Our Final Century?* and the Future of Humanity Institute, which has generated foundational work in existential and global catastrophic risk, was established. It has been followed by the establishment of the Global Catastrophic Risk Institute (2011), the Centre for the Study of Existential Risk (2012), the Global Challenges Foundation (2012), Kyoto University's Global Survivability Institute (2013), the Future of Life Institute (2014), and the Freiburg Law and Governance of Existential Risk Group (2016), among others. Academic courses and workshops covering existential risk have begun at the University of Copenhagen, University of Gothenburg, University of Warwick, and Harvard University.

Early steps have been taken to tackle existential and global catastrophic risks at the policy level. Notable recent examples include a chapter on existential risk in the 2014 United Kingdom Chief Scientific Advisor's annual report (Walport & Craig, 2014), and an existential risk report sponsored by the Finnish Government (Farquhar et al 2017). The World Economic Forum launched an annual Global Risk report in 2007. Generally speaking there has been increased focus on emerging technological risks, as well as on climate change, environmental risks, and resource challenges (e.g. see Collins, A. et al, 2018). Several academic groups, non-profits and policy think tanks, including the Garrick Institute for Risk Sciences, the Bulletin of Atomic Scientists and the Nuclear Threat Initiative, have broadened their scope to consider wider categories of global catastrophic risk.

Additionally, there are separate, established research communities and centres concerned with particular global and existential risks, and communities focused on disaster response and resilience, risk analysis, risk communication, and risk governance. This expertise is a vital resource for existential risk research, required to understand which risks have the potential to be global in scope, placing individual threats within a global risk context, and understanding potential opportunities for intervention. However, there is a need for these connections to be strengthened, and for insights and expertise to be shared more extensively.

The desire to broaden the scope of research in catastrophic and existential risk, and further its connections with other areas of expertise, was a key motivation for both the first Cambridge Conference on Catastrophic Risk and this special issue, brings together many papers originally presented at that conference. Below, we'll introduce the papers collected within, and in doing so draw together some common themes and lessons.

2. Conceptualising and classifying catastrophic risk

Views on how catastrophic and existential risks should be conceptualized differ. Existing research typically focuses on discrete 'knock out' events and the direct consequences of these

events. Examples include asteroid impacts, supervolcanoes, pandemic outbreaks, the possibly catastrophic consequences of future physics experiments, and the development of artificial superintelligence.

However, several papers argue that a narrow focus on 'knock outs' misses possible sources of existential catastrophe. Liu, Lauta & Maas (2018) emphasize *boring apocalypses* wherein our civilization ends not with a bang, but a drawn-out whimper: "There are also many other, slower and more intertwined ways in which the world might collapse, without being hit by spectacular hazards" (7). Similarly, Kareiva & Carranza (2018) stress our ignorance of how ecosystems collapse, "it is our ignorance about the dynamics of complex, tightly coupled, human-natural systems with interacting stresses and feedback loops that is the source of our greatest environmentally-dictated existential risk" (3). Indeed, research on mass extinctions suggests that reductions in the resilience of communities must have already occurred for apparent knock-outs (such as asteroid strikes or massive volcanism) to trigger existential catastrophes (see for instance Mitchell, Roopnarine and Angielczyk 2012).

This recognition of more subtle influences on existential and catastrophic risk has two upshots. First, it highlights the need to not only characterise individual shocks, but to better understand how these shocks, and more subtle stresses, affect the stability of the systems in which they occur. Second, it suggests a widening of the scope of existential risk to include smaller hazards (Liu, Lauta & Mass 2018).

Understanding existential risk at the systems-level requires systematic forms of categorization. The question of how best to categorize risks turns in part on our purpose in doing so. Avin et al (2018) develop a systematization aimed at the study and analysis of existential risks. By analysing risks from the perspective of systemic failure points and spread mechanisms, they sidestep some of our ignorance by identifying both weakness in the resilience of our systems, and points of overlap in pathways of impacts associated with the risks. They also emphasize the necessity of a multidisciplinary approach: "compiling a comprehensive list of plausible GCR [global catastrophic risk] scenarios requires exploring the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios that are typically addressed by single disciplines" (2). In communicative contexts, different systematization may be called for. Turchin & Denkenberger (2018) systematize existential and global risks in order to communicate their dangers to policy makers. Such a tool requires abstraction from the details; the authors use a colour-coding system based on severity and probability. These alternative approaches to categorising and communicating risks are likely to be highly complementary.

3. Catastrophic risk: communication and responsibility

Turchin & Denkenberger's focus highlights the critical role communication within public and policy spheres play in shaping existential and catastrophic risk research: a common theme of the special issue.

Kareiva & Carranza (2018) highlight challenges relating to the presentation of environmental catastrophes to the public:

> … popular culture regarding environmental catastrophes is important because it may lead society to look in the wrong places for existential risks. Movies constantly ascribe disasters to corporate greed and a flawed society. This narrative of either the "villain", or the need for social reform is standard movie fare. No big time Hollywood movie has

> ascribed an environmental disaster to ignorance about ecosystem dynamics or surprises. (10)

Currie (2018), too, considers how public and policy conceptions of emerging technology can affect technology's development. Reflecting on the wider repercussions of SPICE's cancelled geoengineering experiment, he argues that scientific governance cannot be divorced from social, political and technological consequences.

> … the consequences of governance measures should be considered in the context of science's 'ecosystem': the incentives and drivers which shape research… technological research is not separate from the social, economic and ecological consequences of its use. As our tech becomes increasingly powerful, it becomes increasingly pressing to re-evaluate both how science is governed, and the role of the 'scientist' itself. (10)

Jo Husband's (2018) discussion of bio-security challenges is also sensitive to science in its social context. She is interested in tackling 'dual-use' research, research which despite its potential benefit might be misapplied or repurposed for harm. Husband argues that the issue should be framed in terms of 'responsible science'. On this view, as opposed to employing explicit legal frameworks, or taking worries concerning dual-use to be outside of the scope of science, the "potential risks of dual use research are framed as one part of the general social responsibility of science, along with other ethical issues… [this] *makes scientists part of the solution, not part of the problem*" (6). Just as avoiding plagiarism is standardly the scientist's responsibility, in many contexts consideration of research's potential for dual use should be too.

A focus on the role of communication and framing also looms large in Claire Craig's (2018) reflections on risk communication. She provides an overview of the official British system of communication, drawing on the UK government's reaction to the Fukushima disaster. As she says, "Choice of language is at the heart of the challenge of engaging policy-makers and publics with risk, both mundane and extreme" (3). Her discussion is sensitive to tensions between human psychology, the aims of policymakers, and the aims of scientists: "politicians are concerned with avoiding the error of having not acted when they should, while scientists may be more concerned with the error of reporting a finding where the finding is not robust" (3). Navigating diverse aims is necessary not only for research into existential risk, but—as Craig shows—putting that research into effective use.

Crowley, Shang & Dando (2018) are also concerned with how to influence and communicate, but the audience here shifts from scientists and the public to policy-makers. They describe work on emerging scientific developments which present new challenges for the chemical weapons convention, specifically focusing on the possible misuse of advances in neuroscience to develop novel forms of so-called 'nonlethal chemical incapacitating agents'. The paper highlights the fast pace of technological change, and the troubling limitations this puts on existing governance structures: "Such selective manipulations of the human machine may not have been possible then, but they may be implemented in the future if we are not careful" (5). Crowley et al (2018) and Binder (2018) both highlight that managing catastrophic risks has a legal dimension, and that legal mechanisms provide powerful levers for mitigating and avoiding tragedy. Where Crowley, Shang & Dando consider international governance, Binder considers the increasing use of criminal prosecution in disasters. This reflects a shifting conception of responsibility vis-à-vis disasters across public, legal and scientific spheres, and highlights the importance of developing dynamic governance from the local to the international level.

4. Confronting the limits of our knowledge

One central reason for adopting a dynamic and flexible approach to global catastrophic risk governance and response is that the second order effects of many risks remain poorly understood. This is a common theme in the special issue. John Halstead (2018) considers whether the end result of geoengineering research would be a world with increased or decreased existential risk. He draws on Hilary Greaves' notion of 'complex cluelessness' (2016). Such cluelessness manifests in circumstances where, concerning two future pathways, "there is a plausible case to be made in either direction and it is unclear how to weigh the competing considerations" (11). Kareiva and Carranza's paper explores a related concern: "What worries us is our ecological ignorance—especially our ignorance regarding positive feedbacks and interactions among stresses that amplify environmental perturbations in ways that surprise us" (9-10).

We are not only in a situation where it is difficult to ascertain the probabilities of certain identified risks, we are also not sure what the nature and dynamics of the risks might be. As we've mentioned, one source of this ignorance is the unprecedented nature of our situation, as Halstead emphasizes for climate change: "Warming of >6°C would bring temperatures that are far outside the experience of homo sapiens and its immediate ancestors, constituting a radical departure from the Holocene climate in which human civilization has thrived" (Halstead 2018, 3). Shifting into such unprecedented circumstances requires a systematic but creative approach to understanding existential and catastrophic risk (Currie, forthcoming).

This systematic creativity is present in Denkenberger & Wade's (2018) discussion of the mitigation of supervolcanic eruptions. Many of the interventions evaluated are ambitious and untested, but a systematic assessment of the space of possible approaches is essential for such rare disasters. What risks, which sacrifices, and what mitigating actions are we willing to put up with, consider or condone? This leads us to consider how an existential risk perspective might change our approach to particular issues, and what lessons might be exportable from outside of the current boundaries of existential and catastrophic risk (porous as they may be) into those communities.

5. A broader scope and methodological toolkit

John Halstead's analysis of geoengineering from an existential risk perspective is an example of how considering questions from vis-à-vis their relevance to species-level threats can shed new light on existing questions. Other papers in the special issue demonstrate the value of taking lessons from other disciplines in order to understand catastrophic risk.

Kim & Song (2018) look at how urban climate resilience has developed in light of a longer-term perspective: "in recent years, exogenous changes and risks, such as depopulation, aging society, climate change, and disasters, have been emerging, and thus it is important for urban planning to reflect the various components of the urban system" (2). The use and adaptation of novel methodologies, such as DELPHI techniques (Kim & Song 2018) and other horizon scanning exercises (Aicardi et al 2018), has borne fruit in technological forecasting and urban planning. Even games can be useful tools for communication and debate (Craig 2018)[2]. These examples

[2] In addition to more traditional board games like *Pandemic*, some recent computer games have been designed specifically to raise issues about existential risk, including Frank Lantz's *Universal Paperclips*, and a mod for *Civilization V* which adds artificial intelligence developed under Shahar Avin's guidance.

highlight the extent to which tools and methodologies developed in related fields and for related challenges may prove valuable for the study of existential and catastrophic risk.

An emerging field grappling with uncertainty should be open to critiques of community focus, as illustrated by Aicardi et al (2018)'s discussion of the risks considered by the Human Brain Project:

> The growing amount of speculation about general Artificial Intelligence (with some suggesting that we will soon reach the 'singularity', the point at which machine intelligence overtakes human intelligence) is somewhat premature, and diverts our attention from more pressing social and ethical issues arising in connection to the proliferation and rapidly-growing efficiency of not-so-intelligent machine intelligence—of specialized artificial intelligence, as opposed to general artificial intelligence. (3)

Too strong a focus on certain kinds of risk potentially blinds us to others (a warning echoed in Kareiva & Carranza and Liu, Lauta & Maas). This is an important lesson. However, consideration of different types of risk needn't be a zero-sum game. A thriving ecosystem of existential and catastrophic risk research can encompass slow-moving threats and cascading failures, as well as abrupt catastrophic events. It can also encompass well-defined near-term threats and longer-term, more theoretical scenarios. Moreover, engaging with the near-term development of emerging technologies may aid us in avoiding longer-term threats. It encourages a culture of democratic reflection on the development of technology, and drives us to ensure that both the right expertise, and an appropriate diversity of voices and stakeholders, are involved at every stage of the process. Embedding responsibility and accountability in the development of transformative technologies is likely to increase the probability of these technologies achieving their potential for global good, while also increasing the likelihood that potential catastrophic consequences are identified, given due attention, and averted.

A critical place for such democratic reflection is the development of government institutions. Jones, O'Brien & Ryan (2018) survey initiatives to represent the interests of future generations in the governance of a range of nations, and adapt these to the UK context: "the issue of representing the rights and interests of future generations is not well articulated (if at all) in the UK political context. Simple awareness of these issues is an essential step towards their having an impact upon decision making" (2). The paper recommends the establishment of an All-Party Parliamentary Group on Future Generations, which the authors have since launched in January 2018. This body establishes existential risk and the rights of future generations as issues of cross-party interest, and may provide a useful stepping stone to the eventual institutionalisation of intergenerational justice in the UK Parliament.

6. Shaping the future from the perspective of existential and catastrophic risk

A broader conceptualisation of the challenges existential and catastrophic risk bring leads naturally to reflecting on how science and society should be organised in a world with powerful technological and scientific capabilities. Taking existential risk seriously requires fostering the kinds of institutions that will encourage a longer-term perspective, and working to avoid sleep-walking into catastrophic futures. It requires a much more dynamic conception of technological governance. As Craig puts it "society does not have to be the victim of technological determinism or of existing views of human nature" (7).

It also poses the question: how broad should the scope of our interest be? Given that further catastrophic and existential risks remain to be discovered, and that engaging with both known

and unknown risks requires dynamic policy and governance at multiple levels and on multiple timescales, it may be difficult to set strict bounds on the remit of existential risk as a field.

We think there is a clear need for specific work focused on discrete events that might threaten humanity, as well as on the slower-moving or cascading sequences of events that could lead to the same result, and on the methodologies that allow us to predict and mitigate these risks. But we can also conceive of existential risk as a framework from which to analyse a range of developments and challenges. This would include assessing new technological developments via their influence on various aspects of global risk, as illustrated by Halstead's analysis of geoengineering. It would also include assessing the suitability of our scientific processes and our governmental institutions for a future in which we must be prepared to confront risks of this scale. Drawing on Jones, O'Brien & Ryan's institutionalized regard for future generations, as well as Husband's widening of scientific responsibility, we might conclude that a concern for the long-term, the catastrophic, and the existential ought to be a standard part of our technological and institutional development, incorporated into the portfolio of principles that should guide our progress.

The future of research into catastrophic and existential risk, then, involves increasing multi-disciplinarity, creativity and shifting to an applied footing. As we move from studying to mitigating existential and global catastrophic risks, there are many processes we need a deeper understanding of. Excellent work has been done to characterize discrete events that would result in a global catastrophe, although more work certainly remains to be done. However, many causes of global catastrophe may be slower-moving, or take the form of sequences of events that cascade into a global catastrophe. Understanding these more complex risks requires a combination of insights from different domains of risk and intellectual disciplines, as well as the development of specific methodologies. The difficulty of engaging with these complex risks may mean that they are understudied within this community compared to 'big bang' events.

We also need to combine analysis and intervention. This requires understanding how to engage with policymakers, technology leaders, and the broader public on the necessary steps. Here too there is much to learn both from the successes and challenges of communities focused on individual global threats, as well as the lessons learned from engagement at different levels of risk. Papers within the special issue shed light on each of these challenges, drawing out crucial insights for those studying existential risks going forward.

We would argue that *the perspective of existential and catastrophic risk* should be built into how we shape the future. If we wish to continue to reap the benefits of human civilizational progress on our fragile blue planet, while avoiding the worst potential consequences of our unique capabilities as a species, such a perspective is both necessary and urgent.

Bibliography

Papers in the Special Issue

Aicardi, C., Fothergill, B. T., Rainey, S., Stahl, B., & Harris, E. (2018). Accompanying technology development in the Human Brain Project: From foresight to ethics management. Futures.

Avin, S., Wintle, B. C., Weitzdörfer, J., Ó hÉigeartaigh, S. S., Sutherland, W. J., & Rees, M. J. (2018). Classifying global catastrophic risks. Futures.

Binder, D. (2018). The findings of an empirical study of the application of criminal law in non-terrorist disasters and tragedies. Futures.

Craig, C. (2018). Risk management in a policy environment: The particular challenges associated with extreme risks. Futures.

Crowley, M., Shang, L., & Dando, M. (2018). Preserving the norm against chemical weapons: A civil society initiative for the 2018 4th review conference of the chemical weapons convention. Futures.

Currie, A. (2018). Geoengineering tensions. Futures.

Denkenberger, D. C., & Blair, R. W. (2018). Interventions that May Prevent or Mollify Supervolcanic Eruptions. Futures.

Donghyun Kim, Seul-Ki Song. (2018). Measuring changes in urban functional capacity for climate resilience: Perspectives from Korea. Futures.

Halstead, J. (2018). Stratospheric aerosol injection research and existential risk. Futures.

Husbands, J. L. (2018). The challenge of framing for efforts to mitigate the risks of "dual use" research in the life sciences. Futures.

Jones, N., O'Brien, M., & Ryan, T. (2018). Representation of future generations in United Kingdom policy-making. Futures.

Kareiva, P., & Carranza, V. (2018). Existential Risk due to Ecosystem Collapse: Nature Strikes Back. Futures.

Liu, H. Y., Lauta, K. C., & Maas, M. M. (2018). Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research. Futures.

Turchin, A., & Denkenberger, D. (2018). Global catastrophic and existential risks communication scale. Futures.

Other Papers Cited.

Ambrose, S. H. (1998). Late Pleistocene human population bottlenecks, volcanic winter, and differentiation of modern humans. Journal of Human Evolution, 34(6), 623-651.

Bostrom, N. (2013). Existential risk prevention as global priority. Global Policy, 4(1), 15-31.

Collins, A. et al (2018). The Global Risks Report 2018 13th Edition. World Economic Forum.

Currie, A. (forthcoming). Existential Risk, Creativity & Well-Adapted Science. Studies in the History and Philosophy of Science.

Farquhar, S., Halstead, J., Cotton-Barratt, O., Schulbert, S., Belfield, H., & Snyder-Beattie, A. (2017). Existential Risk: Diplomacy and Governance. Global Priorities Project.

Mitchell, J. S., Roopnarine, P. D., & Angielczyk, K. D. (2012). Late Cretaceous restructuring of terrestrial communities facilitated the end-Cretaceous mass extinction in North America. Proceedings of the National Academy of Sciences, 109(46), 18857-18861.

ÓhÉigeartaigh, S. S. (2017). The State of Research in Existential Risk. Proceedings of the First International Colloquium on Catastrophic and Existential Risk. The B. John Garrick Institute for the Risk Sciences, UCLA.

Rees, M. (2003). Our final century? William Heinemann Ltd

Walport, M., & Craig, C. (2014). Innovation: Managing risk, not avoiding it. Annual Report of the Government Chief Scientific Adviser.

Williams, M. (2012). Did the 73 ka Toba super-eruption have an enduring effect? Insights from genetics, prehistoric archaeology, pollen analysis, stable isotope geochemistry, geomorphology, ice cores, and climate models. Quaternary International, 269, 87-93.