Centre for the Study of Existential Risk
Six Month Review: November 2017 – May 2018

We have just prepared a Six Month Report for our Management Committee. This is a public version of that Report. We send short monthly updates in our newsletter – subscribe here.

Contents

1. Overview 1
2. Publications 2
3. Workshops and Public Lectures 5
4. Media Coverage 6
5. Policy Engagement 7
6. Academic Engagement 7
7. Upcoming activities 8

1. Overview

The Centre for the Study of Existential Risk (CSER) is dedicated to the study and mitigation of risks that could lead to civilizational collapse or human extinction. We are an interdisciplinary research centre within the University of Cambridge that studies existential risk, develops collaborative strategies to reduce them, and fosters a global community of academics, technologists and policy-makers working to tackle these risks. Our research focuses on Global Catastrophic Biological Risks, Extreme Risks and the Global Environment, Risks from Artificial Intelligence, and Managing Extreme Technological Risks.

CSER staff, along with our Management Committee and other senior advisors, have continued a high level of research and engagement over the last six months:

- Publication of seventeen papers and a book – some award-winning, one in *Nature*;
- Worldwide media coverage and extensive policy-maker interest for landmark publications on: environmental risk and governance; the malicious use of AI; and emerging issues in bioengineering;
- Hosting three workshops, four public lectures with distinguished speakers, and our second Cambridge Conference on Catastrophic Risk;
- Engagement with the public through media coverage and public talks;
- Supporting Cambridge students to launch an All-Party Parliamentary Group for Future Generations in the UK Parliament.

2. Publications include:

A special issue of the journal *Futures*, based on papers presented at our 2016 Conference, and edited by Adrian Currie, will shortly be published. Several of the papers have already been published online, including contributions from the CSER team:

- Classifying Global Catastrophic Risks. *Futures*. **Shahar Avin, Bonnie Wintle, Julius Weitzdörfer, Seán Ó hÉigeartaigh**, William Sutherland, Martin Rees. (2018).
    - "We present a novel classification framework for severe global catastrophic risk scenarios. Extending beyond existing work that identifies individual risk scenarios, we propose analysing global catastrophic risks along three dimensions: the critical systems affected, global spread mechanisms, and prevention and mitigation failures. The classification highlights areas of convergence between risk scenarios, which supports prioritisation of particular research and of policy interventions. It also points to potential knowledge gaps regarding catastrophic risks, and provides an interdisciplinary structure for mapping and tracking the multitude of factors that could contribute to global catastrophic risks."
- Geoengineering Tensions. *Futures*. **Adrian Currie**. (2018).
    - "There has been much discussion of the moral, legal and prudential implications of geoengineering, and of governance structures for both the research and deployment of such technologies. However, insufficient attention has been paid to how such measures might affect geoengineering in terms of the incentive structures which underwrite scientific progress. There is a tension between the features that make science productive, and the need to govern geoengineering research, which has thus far gone underappreciated. I emphasize how geoengineering research requires governance which reaches beyond science's traditional boundaries, and moreover requires knowledge which itself reaches beyond what we traditionally expect scientists to know about. How we govern emerging technologies should be sensitive to the incentive structures which drive science."

Risks and responsible development of artificial intelligence:

- The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. *arXiv.* Miles Brundage, **Shahar Avin** (joint lead authors), Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, **Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield**, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, Dario Amodei. (2018).
    - "This report surveys the landscape of potential security threats from malicious uses of AI, and proposes ways to better forecast, prevent, and mitigate these threats. After analyzing the ways in which AI may influence the threat landscape in the digital, physical, and political domains, we make four high-level recommendations for AI researchers and other stakeholders. We also suggest several promising areas for further research that could expand the portfolio of defenses, or make attacks less effective or harder to

execute. Finally, we discuss, but do not conclusively resolve, the long-term equilibrium of attackers and defenders."

- An AI Race for Strategic Advantage: Rhetoric and Risks. AAAI/ACM Digital Libraries. Stephen Cave, **Seán Ó hÉigeartaigh**. (2018).
    - Jointly won Best Paper Award.
    - "The rhetoric of the race for strategic advantage is increasingly being used with regard to the development of artificial intelligence (AI), sometimes in a military context, but also more broadly. This rhetoric also reflects real shifts in strategy, as industry research groups compete for a limited pool of talented researchers, and nation states such as China announce ambitious goals for global leadership in AI. This paper assesses the potential risks of the AI race narrative and of an actual competitive race to develop AI, such as incentivising corner-cutting on safety and governance, or increasing the risk of conflict. It explores the role of the research community in responding to these risks. And it briefly explores alternative ways in which the rush to develop powerful AI could be framed so as instead to foster collaboration and responsible progress."

Horizon-scanning and expert elicitation:

- Point of View: A transatlantic perspective on 20 emerging issues in biological engineering. *eLIFE.* **Bonnie Wintle**, **Catherine Rhodes**, **Seán Ó hÉigeartaigh**, Christian R. Boehm, William Sutherland, Robert Doubleday, Jennifer C Molloy, Piers Millett, Laura Adam, Rainer Breitling, Rob Carlson, Rocco Casagrande, Malcolm Dando, Eric Drexler, Brett Edwards, Tol Ellis, Nicholas G Evans, Richard Hammond, Jim Haseloff, Linda Kahl, Todd Kuiken, Benjamin R Lichman, Colette A Matthewman, Johnathan A Napier, Nicola J Patron, Edward Perello, Philip Shapira, Joyce Tait, Eriko Takano. (2017).
    - "Advances in biological engineering are likely to have substantial impacts on global society. To explore these potential impacts we ran a horizon scanning exercise to capture a range of perspectives on the opportunities and risks presented by biological engineering. We first identified 70 potential issues, and then used an iterative process to prioritise 20 issues that we considered to be emerging, to have potential global impact, and to be relatively unknown outside the field of biological engineering. The issues identified may be of interest to researchers, businesses and policy makers in sectors such as health, energy, agriculture and the environment."
- The Value of Performance Weights and Discussion in Aggregated Expert Judgments. *Risk Analysis*. Anca M. Hanea, Marissa F. McBride, Mark A. Burgman, **Bonnie Wintle**. (2018).
    - "In risky situations characterized by imminent decisions, scarce resources, and insufficient data, policymakers rely on experts to estimate model parameters and their associated uncertainties. Different elicitation and aggregation methods can vary substantially in their efficacy and robustness. While it is generally agreed that biases in expert judgments can be mitigated using structured elicitations involving groups rather than individuals, there is still some disagreement about how to best elicit and aggregate judgments. This mostly concerns the merits of using performance-based weighting schemes to combine judgments of different individuals (rather than assigning equal weights to individual experts), and the way that interaction between experts should be handled. This article

aims to contribute to, and complement, the ongoing discussion on these topics."

Global catastrophic risk research and international governance:

- The State of Research in Existential Risk. Garrick, B.J. (Ed.), *Catastrophic and Existential Risk: Proceedings of the First Colloquium*. (Garrick Institute for the Risk Sciences). **Seán Ó hÉigeartaigh**. (2017).
    - "In the last fifteen years there has been substantial growth in research on existential risk – the category of risks that threaten human extinction, or the permanent and drastic reduction of humanity's future potential. A number of new organisations focused explicitly on existential and global catastrophic risk have been founded in recent years, complementing the long-standing work of existing centres focused on specific risk areas such as nuclear war, biosecurity, climate change and systemic risk. This paper provides a brief overview of the emergence of this new research community, and provides a case study on the community's research on potential risks posed by future developments in artificial intelligence. There exists the opportunity for powerful collaboration between the new approaches and perspectives provided by the existential risk research community, and the expertise and tools developed by the risk sciences for risks of various magnitudes. However, there are a number of key characteristics of existential and global catastrophic risks, such as their magnitude, and their rare or unprecedented nature, that are likely to make them particularly challenging to submit to standard risk analysis, and will require new and specialised approaches."
- Risks and Risk Management in Systems of International Governance. Garrick, B.J. (Ed.), *Catastrophic and Existential Risk: Proceedings of the First Colloquium* (Garrick Institute for the Risk Sciences). **Catherine Rhodes**. (2017).
    - "International governance systems will have an important role in management of existential and global catastrophic risks, which by their nature have global impacts, and require coordinated international responses. This paper outlines some of the main contributions international governance systems can make to risk management and draws out some of the ways in which they can fail and be a source of risk. This motivates work to better understand and analyse international governance systems' capabilities and deficiencies in relation to specific sources of existential and catastrophic risks, and so the paper is framed by an effort to outline steps through which this can be done as a contribution to building broader research agendas on managing these classes of risk."

Governance and biodiversity:

- Successful conservation of global waterbird populations depends on effective governance. *Nature*. **Tatsuya Amano**, Tamás Székely, Brody Sandel, Szabolcs Nagy, Taej Mundkur, Tom Langendoen, Daniel Blanco, Candan U. Soykan, William Sutherland. (2017).
- Does governance play a role in the distribution of invasive alien species? *Ecology and Evolution*. Thomas Evans, Philine zu Ermgassen, **Tatsuya Amano**, Kelvin S.-H. Peh. (2018).

- Governance explains variation in national responses to the biodiversity crisis. *Environmental Conservation*. Zachary Baynham-Herd, **Tatsuya Amano**, William Sutherland, Paul F. Donald. (2018).

Issues in decision theory relevant to advanced artificial intelligence:

- Heart of DARCness. *Australasian Journal of Philosophy*. **Yang Liu**, **Huw Price**. (2018).
- A simpler and more realistic subjective decision theory. *Synthese.* Haim Gaifman, **Yang Liu**. (2017).

3. Workshops and Public Lectures

Our events over the last few months have included:

- October: Public Lecture by **Professor Max Tegmark**, on 'Life 3.0: Being human in an age of artificial intelligence'. Prof. Tegmark is an MIT professor and co-founder of the Future of Life Institute. Video.
    - Followed by an award ceremony. **Vasili Arkhipov** single-handedly prevented nuclear war during the Cuban Missile Crisis. 55 years later, his family received an award in his honour in a touching ceremony. Video.

- November: Public Lecture by **Dr Laura Kahn** on 'Meat, monkeys and mosquitoes: a One Health perspective on emerging disease'. Dr Kahn is a Research Scholar at Princeton University's Woodrow Wilson School and a columnist for the Bulletin of the Atomic Scientists. Video.

- December: **Nuclear "Error and Terror" Response and Recovery Policies** Workshop (led by Dr Julius Weitzdörfer). It explored challenges of and solutions to response and recovery following a deliberate or accidental release of radiological and nuclear material. It was co-organised with the UK Government Department for Environment, Food and Rural Affairs (DEFRA). Overview.

- January: **Scientific Governance at the Ground Level** Workshop (led by Dr Adrian Currie). It considered two related questions about the governance of emerging technologies. First, what challenges, limitations, pitfalls and successes emerge from current efforts? Second, how might those challenges be overcome or mitigated, perhaps with alternative models of governance? Overview.
    - Followed by a Public Lecture by **Professor Sabina Leonelli** on 'How to (Re)Use Big Data'. Prof. Leonelli is Professor of Philosophy and History of Science at the University of Exeter, and Co-Director of the Exeter Centre for the Study of the Life Sciences. Video.

- February: **Modelling Societal Collapse** Workshop (led by Haydn Belfield). It explored whether it is possible to compare societal collapses and build predictive models, or whether each collapse is "unhappy in its own way". It was co-organised with the Cambridge Conservation Initiative.

  - Followed by a Public Lecture by **Professor Jared Diamond** on 'National crises, viewed in the light of personal crises'. Prof. Diamond is Professor of Geography at UCLA, and the Pulitzer-Prize winning author of *Guns, Germs, and Steel* and *Collapse*.

- April: **Cambridge Conference on Catastrophic Risk 2018**. Our second major conference discussed recent developments in the field, and specific challenges of existential risk research. It built on recommendations from our **2016 conference**, and contributed to the further development of the community working in this field. It featured keynote talks from leading thinkers including Tamsin Edwards (climate modelling), Karin Kuhlemann (overpopulation) and Peter Ho (technology policy). Keynote videos forthcoming.

4. Media Coverage

- *The Malicious Use of AI* report received **worldwide press coverage**, including on the BBC, Al Jazeera, NYT, Wired, Guardian, Financial Times, etc. A UK minister, the Commander of the Australian Defence College, and the former President of AAAI, praised it. We published opinion pieces in **Wired**, the **Telegraph** and **People's Daily** (the Chinese Communist Party's official newspaper).
- Tatsuya Amano's *Nature* paper on the link between effective governance and biodiversity conservation (based on analysis of waterbirds in wetlands) was **covered around the world**, for example in France's Le Monde, and Spain's ABC.
- Our bioengineering horizon-scan paper was featured on BBC Inside Science and the eLIFE podcast.

We're able to reach far more people with our research:

- Since our new site launched in Aug 2017, we've had 33,909 visitors.
- 6,211 newsletter subscribers, up from 4,863 in Oct 2016.
- Facebook followers have roughly **tripled** since Dec 2016, from 627 to 1,874.
- Twitter followers have over **quintupled** since Dec 2016, from 778 to 4,010.

- Dr Rhodes gave a **TEDx talk**: Bringing Existential Risk Home.
- Dr Shahar Avin released a 'mod' for the **popular strategy video game** Civilization V, which received extensive media coverage.
- Dr Currie wrote an article in Aeon, Does science need mavericks?, that was **shared over 700 times**.
- Lord Rees was interviewed by the **Financial Times** on Tech Tonic podcast and by 'Vision'.
- Dr Simon Beard is an **AHRC/BBC 'Next Generation Thinker'**. He wrote articles on AI and justice, lessons from GMOs, geoengineering, and overpopulation.

- Dr Weitzdörfer was quoted in a Telegraph article on our 'Black Sky' research.
- Dr Ó hÉigeartaigh was interviewed in Spain's El Periodico, wrote for Cambridge University's Research Horizons, and was quoted in Verge on the Doomsday Clock.
- CSER staff and co-founders featured in a video giving an overview of our work.

5. Policy Engagement

- Partha Dasgupta participated in a conference at **the Vatican on the link between climate change and public health**, Health of People and Planet: Our Responsibility. It is the third in a series of workshops with the Pontifical Academy of Sciences. The first in 2015 influenced the landmark Papal Encyclical on Climate Change. The second in 2017 on 'Biological Extinction' is being turned into a book for Cambridge University Press.

- MPs, Peers, academics and industry leaders launched the **All-Party Parliamentary Group for Future Generations** in January in Parliament. The 'APPG', created by Cambridge students, helped by CSER staff, and chaired by Daniel Zeichner MP, aims to challenge political short-termism and raise the profile of the long-term interests of future generations. More.

- We were invited to present our bioengineering horizon-scan paper at the **2017 Meeting of States Parties to the Biological Weapons Convention** (Dr Lalitha Sundaram presented and spoke at the Meeting's press conference), and to the **Science Advisory Board of the Organisation for the Prohibition of Chemical Weapons** (Dr Bonnie Wintle presented).

- We have had extensive discussions with policy-makers on the findings of *The Malicious Use of AI*. We submitted evidence to the House of Lords Select Committee, and presented findings to the Committee on their visit to Cambridge. The UK has staked out a **leadership position on ethical AI**. Prime Minister Theresa May said she wants the UK to lead the world in deciding how to deploy AI in a safe and ethical manner, and the UK will have the world's first national advisory body for AI.

- Sean O hEigeartaigh gave a keynote interview on emerging threats on the intersection of AI and cybersecurity at **Thomson Reuters' risk summit**, and spoke on AI and sustainability at **The Crowd** – an industry sustainability event.

- CSER researchers continued meetings with top UK civil servants as part of the policy fellows program organized by the Centre for Science and Policy (**CSaP**).

6. Academic Engagement

- CSER signed a memorandum of understanding on an **academic partnership** with Kyoto University's Graduate School of Advanced Integrated Studies in Human Survivability (**GSAIS**, or "Shishu-Kan").

- Dr Rhodes participated in the **Gothenburg Centre for Advanced Studies programme** on Existential Risk to Humanity, including a talk at the opening conference. Video. She also presented at the **Nanotechnology Doctoral Training Centre**, the Interacademy Partnership International **Meeting on Security Implications of Genome Editing Technology**, Public Policy Exchange Symposium, and 2017 **EUSynBioS** symposium.

- CSER co-organised the AI & Society Symposium and Beneficial AI Tokyo, attended by hundreds of academics, technologists and policy-makers.

- Dr Ó hÉigeartaigh, Dr Yang Liu, and colleagues from the Centre for the Future of Intelligence led a delegation from Cambridge to a **high-level workshop in China**.

- Haydn Belfield and Dr Avin led a workshop and lectured at the **Effective Altruism Global London** conference, encouraging students to start careers in existential risk research.

- Dr Weitzdörfer visited Sizewell B nuclear power station, and presented at the Third Northern European **Conference on Emergency and Disaster Studies** and 'Japanese Studies and the Environmental Humanities' in Oxford.

- Dr Liu gave talks at **Beijing Normal University** and the **Institute of Logic and Intelligence**, Southwest University in China, and the 6th International **Conference on Logic, Rationality and Interaction** in Japan.

7. Upcoming activities

- **Visiting researchers:** We will have several visitors over the next few months, including STS and security expert Sam Weiss-Evans, environmental management expert Asaf Tzachor, Rush Stewart from the Munich Centre for Mathematical Philosophy, and Frank Roevekamp, working on insuring against hidden existential risks.

- **Workshop plans under development:** Details are under development, but topics include: a biosecurity workshop, a catastrophic food security risk workshop, and several workshops building on issues raised in the *Malicious Use of AI* report.

- **San Francisco:** Avin and Belfield will present at June's Effective Altruism Global.

- **Munich:** Yang Liu will be co-organising the 2nd conference on Decision Theory and AI in Munich from 26-28th July.