Centre for the Study of Existential Risk Six Month Report: November 2018 – April 2019

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

Contents

1. Overview

The Centre for the Study of Existential Risk (CSER) is an interdisciplinary research centre within the University of Cambridge dedicated to the study and mitigation of risks that could lead to civilizational collapse or human extinction. We study existential risk, develop collaborative strategies to reduce them, and foster 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.

Our last Management Board Report was in October 2018. Over the last five months, we have continued to advance existential risk research and grow the field. Highlights include:

- Publication of *Extremes* book, seven papers in venues like *Nature Machine Intelligence,* and a Special Issue.
- Engagement with global policymakers and industry-leaders at conferences, and in one-on-one meetings.
- Announcement that Prof. Dasgupta will lead the UK Government Global Review of the Economics of Biodiversity.
- Submission of advice to key US, UN and EU advisory bodies.
- Hosting of several expert workshops, helping us to *inter alia* encourage leading machine learning researchers to produce over 20 AI safety papers.
- Welcomed new research staff and visitors.
- Produced a report on business school rankings, contributing to the two leading business school rankers reviewing their methodology.
- Public engagement through media coverage and the exhibition 'Ground Zero Earth'.

2. Policy Engagement:

We have had the opportunity to speak directly with policymakers and institutions across the world who are grappling with the difficult and novel challenge of how to unlock the socially beneficial aspects of new technologies while mitigating their risks. Through advice and discussions, we have the opportunity to reframe the policy debate and to hopefully shape the trajectory of these technologies themselves.

- Prof. Sir Partha Dasgupta, the Chair of CSER's Management Board, will lead the UK Government comprehensive **global review of the link between biodiversity and economic growth**. The aim is to "explore ways to enhance the natural environment and deliver prosperity". The announcement was made by the Chancellor of the Exchequer in the Spring Statement.

- Submitted advice to the **UN High-level Panel on Digital Cooperation** (Luke Kemp, Haydn Belfield, Seán Ó hÉigeartaigh, Zoe Cremer). CSER and FHI researchers laid out the challenges posed by AI and offered some options for the global, international governance of AI. The Secretary-General established the Panel, which Melinda Gates and Jack Ma co-chair. The Panel chose this advice as one of five from over 150 submissions to be highlighted at a 'virtual town hall'. The advice may influence global policy-makers and help set the agenda. Read Advice.

- Submitted advice to the **EU High-Level Expert Group on Artificial Intelligence.** Haydn Belfield and Shahar Avin respond to the Draft Ethics Guidelines for Trustworthy AI, drawing attention to the recommendations in our The Malicious Use of Artificial Intelligence report. This helped influence the EU's Ethics Guidelines, affecting behaviour across the Europe. Read Advice.

- The **All-Party Parliamentary Group for Future Generations, set** up by Cambridge students mentored by CSER researchers, held an event on Global Pandemics: Is the UK Prepared? in Parliament in November 2019, continuing our engagement with UK parliamentarians on existential risk topics. Speakers: Dr Catherine Rhodes (CSER), Dr Piers Millett (FHI), Professor David Heymann CBE (London School of Hygiene and Tropical Medicine). Report here. The APPG has also recently hired two Coordinators, Sam Hilton and Caroline Baylon.

- Submitted advice to the US Government's **Bureau of Industry and Security** on "Review of Controls on Certain Emerging Technologies" (Sam Weiss Evans). The Bureau is the part of the US government that controls the US export control regime. Read Advice.

- Shahar Avin advised the **UK government**, including the Centre for Data Ethics and Innovation (the UK's national AI advisory body). This kind of engagement is crucial to ensuring research papers actually have an impact, and do not just gather dust on the shelf.

- Sean Ó hÉigeartaigh was one of 50 experts exclusively invited to participate in the second Global A.I Governance Forum at the **World Government Summit** in Dubai. The Summit is dedicated to shaping the future of governments worldwide.

- CSER researchers attended **invite-only events** on Modern Deterrence (Ditchley Park), and High impact bio-threats (Wilton Park).

- At the **United Nations**, CSER researchers attended the negotiations on Lethal Autonomous Weapons Systems (LAWS) and the Biological Weapons Convention annual meeting of states parties. They also engaged with the United Nations Institute for Disarmament Research (UNIDIR).

- 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**).

3. Academic and Industry Engagement:

As an interdisciplinary research centre within Cambridge University, we seek to grow the academic field of existential risk research, so that it receives the rigorous and detailed attention it deserves. Researchers also continued their extensive and deep collaboration with industry. Extending our links improves our research by exposing us to the cutting edge of industrial R&D, and helps to nudge powerful companies towards more responsible practices.

- Several researchers participated in the **Beneficial AI Puerto Rico** Conference, engaging with industry and academic leaders, and shaping the agenda of the AI risk community for the next two years. Sean Ó hÉigeartaigh and Shahar Avin gave Keynotes. This was the third conference organised by the Future of Life Institute. The first in 2015 produced a research agenda for safe and beneficial AI, endorsed by thousands of researchers. The second in 2017 produced the Asilomar AI Principles.

- **Visiting researchers**: Dr Kai Spiekermann from LSE visited January-March to work on a paper on 'irreversible losses'; Prof Hiski Haukkala, former foreign policy Adviser to the Finnish President; Dr Simona Chiodo and Dr Daniele Chiffi of the €9m Territorial Fragility project at the Politecnico di Milano.

- Sean Ó hÉigeartaigh attended the **Partnership on AI** meeting and contributed to the creation of several AI/AGI safety- and strategy-relevant project proposals with the Safety-critical AI working group.

- Several CSER researchers contributed to the mammoth ***Ethically Aligned Design**, First Edition: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.* It was produced by IEEE, the world's largest technical professional organization. The release culminates a three-year, global iterative process involving thousands of experts.

- Luke Kemp and Shahar Avin participated in a high-level **AI and political security** workshop led by Prof Toni Erskine at the Coral Bell School of Asia Pacific Affairs at the Australian National University.

- Shahar Avin continued running **'war-games'** exploring different possible AI scenarios. He has run over a dozen so far, with some participants from leading AI labs. He aims to explore the realm of possibilities, and educate participants on some of the challenges ahead.

- We continued our support for the student-run **Engineering Safe AI** reading group. The group exposes masters and PhD students to interesting AI safety research, so they consider careers in that area.

- Catherine Rhodes had several meetings with **groups in Washington DC** working on global catastrophic biological risks, governance of dual-use research in the life sciences, and extreme technological risk more broadly.

- Lalitha Sundaram is working with South African groups to boost their capacity in **low cost viral diagnostics**.

- We will partner with the Journal of Science Policy & Governance to produce a **Special Issue** on governance for dual-use technologies. This Special Issue will encourage students to engage with existential risk research and help us identify future talent.

4. Public Engagement:

- Luke Kemp, Lauren Holt, and Simon Beard had articles published on the BBC with over 1.5m views: Are we on the road to civilisation collapse? and What are the biggest threats to humanity?

- We have published 11 videos of talks given at 2018's Cambridge Conference on Catastrophic Risk, our major international conference.

- Lalitha Sundaram and Simon Beard were featured on leading BBC Radio 4 programme *Analysis* on Will humans survive the century?

- Sean Ó hÉigeartaigh and other researchers were featured in a Cambridge University video on life in the age of intelligent machines.

We are able to reach far more people with our research online:

- 14,000 website visitors over the last 90 days.
- 6,602 newsletter subscribers, up from 4,863 in Oct 2016.
- 6,343 Twitter followers.
- 2,208 Facebook followers.

- Simon Beard produced a BBC radio programme on "I love my children but are they the biggest moral mistake I ever made?"

- Catherine Rhodes was interviewed on the Future of Life Institute Podcast about Governing Biotechnology.

- Catherine Rhodes, Des Browne, Bill Sutherland and David Aldridge had a letter published in *Nature* on Brexit threatening biosecurity.

- Lauren Holt, Paul Upchurch and Simon Beard published a *Conversation* article on global systems failure and the extinction of the dinosaurs.

- Lord Martin Rees was interviewed by Christiane Amanpour on CNN and Stephen Sackur on BBCHardtalk, the Economist, Talking Politics, and Canadian national radio, Academia Europaea, and the Guardian (video). He gave keynotes at the Long Now foundation (video), the European Parliament (video) and the House of Lords (video).

5. Recruitment and research team

New Postdoctoral Research Associates:

Dr Ellen Quigley is working on how to address climate change and biodiversity risks through the investment policies and practices of institutional investors. She was previously a CSER Research Affiliate. She also collaborates with the Centre for Endowment Asset Management at the Judge Business School, who jointly fund her work. She recently published the Business School Rankings for the 21st Century report at events in Davos and Shanghai. Four days later, the Financial Times announced a "complete review of their methodology", supported by a letter in the FT signed by two dozen business leaders.

Dr Jess Whittlestone will work on a research project combining foresight and policy/ethics for AI, in collaboration with the Centre for the Future of Intelligence (CFI) where she is a postdoctoral researcher. She is the lead author on a major new report (and paper) *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research*. It surveys the dozen sets of AI principles proposed over the last two years, and suggests that the next step for the field of AI ethics is to explore the tensions that arise as we try to implement principles in practice.

Visiting researchers:

Lord Des Browne, UK Secretary of State for Defence (2006-2008) and Vice Chair of the Nuclear Threat Initiative. Lord Browne is involved with the new Biosecurity Risk Initiative at St Catharine's College (BioRISC), and will be based at CSER for around a day a week.

Phil Torres, visiting March-June. Author of *Morality, Foresight, and Human Flourishing* (2017) and *The End: What Science and Religion Tell Us about the Apocalypse* (2016). He will work on co-authored papers with Simon Beard and on a new book.

Dr Olaf Corry, visiting March-September. Associate Professor in the Department of Political Science, Copenhagen University. With a background in the international politics of climate change, he will be researching solar geoengineering politics.

Rumtin Sepasspour, visiting Spring/Summer (four months). Foreign Policy Adviser in the Australian Prime Minister's Office, he will focus on enhancing CSER researchers' capability to develop policy ideas

Dr Eva Vivalt, visiting June. Assistant Professor (Economics) at Australia National University, PI on Y Combinator Research's basic income study, and founder of AidGrade, a research institute that generates and synthesizes evidence in international development.

6. Expert Workshops and Public Events:

- November, January: **Epistemic Security** Workshop (led by Dr Avin). Part of a series of workshops co-organised with the UK's Alan Turing Institute, looking at the changing threat landscape of information campaigns and propaganda, given current and expected advances in machine learning. The DSTL is the UK Government's main defence laboratory.

- January: **SafeAI 2019** Workshop (led by Dr Ó hÉigeartaigh and colleagues) at the Association for the Advancement of Artificial Intelligence's (AAAI) Conference. AAAI is one of the four most important AI conferences globally. These regular workshops embed safety in the wider field, and provide a publication venue. The workshop featured over 20 cutting-edge papers in AI safety, and encouraged leading AI researchers to publish on AI safety.

- February-March: **Ground Zero Earth Exhibition**. Curated by CSER Research Affiliate Yasmine Rix, held in collaboration with CRASSH. Five artists explored existential risk. The exhibition was held at the Alison Richard Building, home to the Politics and International Studies Department. The exhibition engaged academics and the public in our research. The launch event was featured on BBC Radio. It closed with a 'Rise of the Machines' short film screening. Read overview.

- March: **Extremes Book Launch**. The book, edited by Julius Weitzdörfer and Duncan Needham, draws on the 2017 Darwin College Lecture Series Julius co-organised. It features contributions from Emily Shuckburgh, Nassim Nicholas Taleb, David Runciman, and others. More.

- March 28-31: **Augmented Intelligence Summit**. The Summit brought together a multi-disciplinary group of policy, research, and business leaders to imagine and interact with a simulated model of a positive future for our global society, economy, and politics – through the lens of advanced AI. Dr Avin was on the Steering Committee, delivered a Keynote, and ran a scenario simulation. More.

- 3-5 April: **EiM 2: The second meeting on Ethics in Mathematics**. Dr Maurice Chiodo and Dr Piers Bursill-Hall from the Faculty of Mathematics in Cambridge have been spearheading an effort to teach responsible behaviour and ethical awareness to mathematicians. CSER supported the workshop. More.

- 5-6 April: **Tools for building trust in AI development** (co-led by Shahar Avin) this two-day workshop convened some of the world's top experts in AI, security, and policy to survey existing mechanisms for trust-building in AI and develop a research agenda for designing new ones.

7. Upcoming activities

Three more books will be published this year:

- **Fukushima and the Law** is edited by Julius Weitzdörfer and Kristian Lauta, and draws upon a 2016 workshop Fukushima – Five Years On, which Julius co-organised.

- **Biological Extinction** is edited by Partha Dasgupta, and draws upon the 2017 workshop with the Vatican's Pontifical Academy of Sciences he co-organised.

- **Time and the Generations - population ethics for a diminishing planet** (New York: Columbia University Press), by Partha Dasgupta, based on his Kenneth Arrow Lectures delivered at Columbia University.

Upcoming events:

- 21 May: **Local Government Climate Futures** (led by Simon Beard with Anne Miller).

- 6-7 June: **Evaluating Extreme Technological Risks** workshop (led by Simon Beard).

- 26 June: The Centre for Science and Policy **(CSaP) Conference.** CSER is partnering on a panel at the conference focusing on methods and techniques for forecasting extreme risks.

- 26-27 August: **Decision Theory & the Future of Artificial Intelligence** Workshop (led by Huw Price and Yang Liu). The third workshop in a series bringing together philosophers, decision theorists, and AI researchers to promote research at the nexus of decision theory and AI. Co-organised with the Munich Center for Mathematical Philosophy.

Timing to be confirmed:

- Summer: **Generality and Intelligence: from Biology to AI.** The next in the Cambridge^2 workshop series, co-organised by the MIT-IBM Watson AI Lab and CFI.

- Summer: **Culture of Science - Security and Dual Use** Workshop (led by Dr Evans).

- Summer/Autumn: **Biological Extinction** symposium, around the publication of Sir Partha's book.

- Autumn: **Horizon-Scanning** workshop (led by Dr Kemp).

- April 2020: CSER's next international conference: the **2020 Cambridge Conference on Catastrophic Risk**.

8. Publications

- Needham, D. and **Weitzdörfer, J.** (Eds). (2019). Extremes. Cambridge University Press.
    - Humanity is confronted by and attracted to extremes. Extreme events shape our thinking, feeling, and actions; they echo in our politics, media, literature, and science. We often associate extremes with crises, disasters, and risks to be averted, yet extremes also have the potential to lead us towards new horizons. Featuring essays by leading intellectuals and public figures (like Emily Shuckburgh, Nassim Nicholas Taleb and David Runciman) arising from the 2017 Darwin College Lectures, this volume explores 'extreme' events.

- Cave, S. and **ÓhÉigeartaigh, S.** (2019). Bridging near-and long-term concerns about AI. *Nature Machine Intelligence* 1:5.
    - We were invited to contribute a paper to the first edition of the new Nature journal, *Nature Machine Intelligence*.
    - "Debate about the impacts of AI is often split into two camps, one associated with the near term and the other with the long term. This divide is a mistake — the connections between the two perspectives deserve more attention."

- Häggström, O. and **Rhodes, C.** (2019). Special Issue: Existential risk to humanity. *Foresight.*
    - Häggström, O. and **Rhodes, C.** (2019). Guest Editorial. *Foresight.*
    - "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."

- **Beard, S.** (2019). What Is Unfair about Unequal Brute Luck? An Intergenerational Puzzle. *Philosophia.*
    - "According to Luck egalitarians, fairness requires us to bring it about that nobody is worse off than others where this results from brute bad luck, but not where they choose or deserve to be so. In this paper, I consider one type of brute bad luck that appears paradigmatic of what a Luck Egalitarian ought to be most concerned about, namely that suffered by people who are born to badly off parents and are less well off as a result. However, when we consider what is supposedly unfair about this kind of unequal brute luck, luck

egalitarians face a dilemma. According to the standard account of luck egalitarianism, differential brute luck is unfair because of its effects on the distribution of goods. Yet, where some parents are worse off because they have chosen to be imprudent, it may be impossible to neutralize these effects without creating a distribution that seems at least as unfair. This, I argue, is problematic for luck egalitarianism. I, therefore, explore two alternative views that can avoid this problem. On the first of these, proposed by Shlomi Segall, the distributional effects of unequal brute luck are unfair only when they make a situation more unequal, but not when they make it more equal. On the second, it is the unequal brute luck itself, rather than its distributional effects, that is unfair. I conclude with some considerations in favour of this second view, while accepting that both are valid responses to the problem I describe.”

- **Beard, S.** (2019). Perfectionism and the Repugnant Conclusion. The Journal of Value Inquiry.
    - “The Repugnant Conclusion and its paradoxes pose a significant problem for outcome evaluation. Derek Parfit has suggested that we may be able to resolve this problem by accepting a view he calls ‘Perfectionism’, which gives lexically superior value to ‘the best things in life’. In this paper, I explore perfectionism and its potential to solve this problem. I argue that perfectionism provides neither a sufficient means of avoiding the Repugnant Conclusion nor a full explanation of its repugnance. This is because even lives that are ‘barely worth living’ may contain the best things in life if they also contain sufficient ‘bad things’, such as suffering or frustration. Therefore, perfectionism can only fully explain or avoid the Repugnant Conclusion if combined with other claims, such as that bad things have an asymmetrical value relative to many good things. This combined view faces the objection that any such asymmetry implies Parfit’s 'Ridiculous Conclusion'. However, I argue that perfectionism itself faces very similar objections, and that these are question-begging against both views. Finally, I show how the combined view that I propose not only explains and avoids the Repugnant Conclusion but also allows us to escape many of its paradoxes as well.”

- **Avin, S.** (2018). Mavericks and lotteries. *Studies in History and Philosophy of Science Part A*.
    - “In 2013 the Health Research Council of New Zealand began a stream of funding entitled ‘Explorer Grants’, and in 2017 changes were introduced to the funding mechanisms of the Volkswagen Foundation ‘Experiment!’ and the New Zealand Science for Technological Innovation challenge ‘Seed Projects’. All three funding streams aim at encouraging novel scientific ideas, and all now employ random selection by lottery as part of the grant selection process. The idea of funding science by lottery emerged independently in several corners of academia, including in philosophy of science. This paper reviews the conceptual and institutional landscape in which this policy proposal emerged, how different academic fields presented and supported arguments for the proposal, and how these have been reflected (or not) in actual policy. The paper presents an analytical synthesis of the arguments presented to date, notes how they support each other and shape policy

recommendations in various ways, and where competing arguments highlight the need for further analysis or more data. In addition, it provides lessons for how philosophers of science can engage in shaping science policy, and in particular, highlights the importance of mixing complementary expertise: it takes a (conceptually diverse) village to raise (good) policy."

- **Avin, S**. (2019). Exploring artificial intelligence futures. *Journal of Artificial Intelligence Humanities Vol.2*.
    - "Artificial intelligence technologies are receiving high levels of attention and 'hype', leading to a range of speculation about futures in which such technologies, and their successors, are commonly deployed. By looking at existing AI futures work, this paper surveys and offers an initial categorisation of, several of the tools available for such futures-exploration, in particular those available to humanities scholars, and discusses some of the benefits and limitations of each. While no tools exist to reliably predict the future of artificial intelligence, several tools can help us expand our range of possible futures in order to reduce unexpected surprises, and to create common languages and models that enable constructive conversations about the kinds of futures we would like to occupy or avoid. The paper points at several tools as particularly promising and currently neglected, calling for more work in data-driven, realistic, integrative, and participatory scenario role-plays."

- Lewis, S.C., Perkins-Kirkpatrick, S.E, Althor, G., King, A.D., **Kemp, L** (2019). Assessing contributions of major emitters' Paris-era decisions to future temperature extremes. *Geophysical Research Letters.*
    - "Temperature extremes can damage aspects of human society, infrastructure, and our ecosystems. The frequency, severity, and duration of high temperatures are increasing in some regions and are projected to continue increasing with further global temperature increases as greenhouse gas emissions rise. While the international Paris Agreement aims to limit warming through emissions reduction pledges, none of the major emitters has made commitments that are aligned with limiting warming to 2 °C. In this analysis, we examine the impact of the world's three largest greenhouse gas emitters' (EU, USA, and China) current and future decisions about carbon dioxide emissions on the occurrence of future extreme temperatures. We show that future extremes depend on the emissions decisions made by the major emitters. By implementing stronger climate pledges, major emitters can reduce the frequency of future extremes and their own calculated contributions to these temperature extremes."

- Hernandez Orallo, J., Martinez-Plumed, F., **Avin, S.**, and **ÓhÉigeartaigh, S**. (2019). Surveying Safety-relevant AI Characteristics. *Proceedings of the AAAI Workshop on Artificial Intelligence Safety 2019*.
    - Shortlisted for Best Paper Prize.
    - "The current analysis in the AI safety literature usually combines a risk or safety issue (e.g., interruptibility) with a particular paradigm for an AI agent (e.g., reinforcement learning). However, there is currently no survey of safety-relevant characteristics of AI systems that may reveal neglected areas of research or

suggest to developers what design choices they could make to avoid or minimise certain safety concerns. In this paper, we take a first step towards delivering such a survey, from two angles. The first features AI system characteristics that are already known to be relevant to safety concerns, including internal system characteristics, characteristics relating to the effect of the external environment on the system, and characteristics relating to the effect of the system on the target environment. The second presents a brief survey of a broad range of AI system characteristics that could prove relevant to safety research, including types of interaction, computation, integration, anticipation, supervision, modification, motivation and achievement. This survey enables further work in exploring system characteristics and design choices that affect safety concerns."

- Report: Pitt-Watterson, D. and **Quigley, E.** (2019). Business School Rankings for the 21st Century.
  - "This paper addresses the question of how business schools, and the courses they offer, are evaluated and ranked. The existing benchmarking systems, many of which are administered by well-respected media institutions, appear to have a strong motivational effect for administrators and prospective students alike. Many of the rankings criteria currently in use were developed years or decades ago, and use simple measures such as salary and salary progression. Less emphasis has been placed on what is taught and learned at the schools. This paper argues that, given the influence of the ranking publications, it is time for a review of the way they evaluate business education. What follows is meant to contribute to a fruitful ongoing discussion about the future of business schools in our current century."