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

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 May 2018. Over the last six months we have continued to advance existential risk research and grow the community working in the field:

- Publication of twelve papers on topics including scientific communities and risk, government reactions to disasters, environmental assessment of high-yield farming, decision theory, and theoretical mapping of artificial intelligence;
- Publication of our Special Issue *Futures of Research in Catastrophic and Existential Risk* featuring fifteen papers, many first presented at our 2016 Conference;
- Hosting five expert workshops, helping us to establish/consolidate partnerships with important players such as the Singaporean Government, the UK Defence Science and Technology Laboratory, Munich University, nuclear experts and MIT;

- Policy-maker engagement with UK Parliamentarians and civil servants, and at the United Nations, where we helped lead a track of the *AI for Good* summit series;
- Academic engagement, building the existential risk field by hosting visiting researchers and presenting at leading conferences;
- Industry engagement, tapping into cutting-edge R&D and nudging companies towards responsibility;
- Recruited two new postdoctoral researchers, a new administrator, and a Senior Research Associate: Academic Programme Manager;
- Continued success in fundraising for CSER's next stage;
- Engaging the public through media coverage (including on Newsnight) and two public lectures with distinguished speakers; and
- The release of Lord Martin Rees' new book, On The Future: Prospects for Humanity.

2. Policy and Industry Engagement:

We have had the opportunity to speak directly with policymakers, industry-leaders 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.

- The **All-Party Parliamentary Group for Future Generations** held two events in Parliament. The APPG was set up by Cambridge students mentored by CSER researchers. This continues our engagement of UK parliamentarians on existential risk topics:
  - Black Sky risks and infrastructure resilience. The main speaker for the evening was Lord Toby Harris, UK coordinator for the Electricity Infrastructure Resilience Council. Julius Weitzdoerfer and Dr Beard also spoke. Overview.
  - How do We Make AI Safe for Humans? This event's speakers were Edward Felten**,** former Deputy White House CTO, Joanna Bryson**,** Reader in AI at the University of Bath, Nick Bostrom.**,** Director of the Future of Humanity Institute, and our own Shahar Avin. Overview.
- The **AI for Good Summit series** is the leading United Nations platform for dialogue on AI. As the UN seeks to enhance its capacity to address AI issues, we have been invited to share our research and expertise. In May, a joint CSER/CFI team led one of the Summit's four 'Tracks', on Trust in AI. This meant we were able to directly shape which topics global policy-makers from many countries and UN departments engaged with, and helped set the agenda for the next year. Overview.
- Shahar Avin has had extensive engagement around the major report *The Malicious Use of Artificial Intelligence*, of which he was the joint lead author. He has presented to the UK **Cabinet Office**, the US' Pacific Northwest National Laboratory (**PNNL**), to the **Dutch Embassy**, and at the Stockholm International Peace Research Institute (**SIPRI**) workshop "Mapping the Impact of Machine Learning and Autonomy on Strategic Stability and Nuclear Risk".
- Dr Avin co-wrote the Digital Catapult AI Ethics Framework. **Digital Catapult** is the UK's leading digital technology innovation centre, funded by the Government. The Framework is intended to be used by AI start-ups. The intention is to nudge AI companies at an early stage, when they are more easily influenced.
- We continued our **industry engagement**. 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. Seán Ó hÉigeartaigh and Dr Avin presented to **Arm**, a leading semiconductor and software design company in Cambridge.
- 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 Engagement:

As an interdisciplinary research centre within the University of Cambridge, we seek to grow the academic field of existential risk research, so that this important topic receives the rigorous and detailed attention it deserves.

- **Visiting researchers:** We have had several visitors, including Dr Rush Stewart from the Munich Centre for Mathematical Philosophy, Dr Frank Roevekamp, working on insuring against hidden existential risks, and Prof David Alexander, Professor of Risk & Disaster Reduction at UCL's Institute for Risk & Disaster Reduction.

- Julius Weitzdörfer gave presentations at UCLA for the **Quantifying Global Catastrophic Risks Workshop** at the Garrick Institute for Risk Sciences, and at a **Special Session on Global and catastrophic risks** at the 14th Probabilistic Safety Assessment & Management Conference. He also gave talks on Disaster, Law and Social Justice in Japan at the **New Perspectives in Japanese Law conference** at Harvard Law School and the **East Asia Seminar Series** in Cambridge.

- In Cambridge, Catherine Rhodes and Sam Weiss Evans presented on the responsible governance of synthetic biology governance. Dr Rhodes and Lalitha Sundaram are coordinators of the **OpenIP Models of Emerging Technologies** seminar series.

- Haydn Belfield and Shahar Avin met collaborators and donors in San Francisco in June, and led workshops at the **Effective Altruism Global** conference.

- Dr Avin presented at the first **AI Humanities** conference in Seoul, at the **Deep Learning in Finance** Summit, the **Big Data & Society** conference at London Metropolitan University, a HSBC risk training event at the Judge Business School, and to Cambridge Computer Science Masters students. He also attended the Origins workshop **Artificial Intelligence and Autonomous Weapons Systems** at Arizona State University, attended by former Secretary of Defence William J. Perry.

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

4. Public Engagement:

- Lord Rees' new book on existential risk, 'On the Future' has received a lot of media coverage, including favourable reviews in the Financial Times, Sunday Times, Vanity Fair, Inside Higher Education, the New Statesman; and interviews with Vox, the Sunday Times, the Harvard Gazette, Australian national radio, several podcasts, and the Chicago Tonight TV show.

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

- Since our new site launched in Aug 2017, we've had 53,726 visitors.
- 6,394 newsletter subscribers, up from 4,863 in Oct 2016.
- Facebook followers have tripled since Dec 2016, from 627 to 2,049.
- Twitter followers have sextupled since Dec 2016, from 778 to 5,184.

- Lalitha Sundaram, Simon Beard, Shahar Avin, and Haydn Belfield gave an interview on the Five Best Books on Existential Risks.

- Simon Beard appeared on Newsnight, the BBC's leading current affairs programme, to discuss the new IPCC report, climate change, and existential risk. Video.

- Simon also appeared on BBC Radio with an essay on existential risk and Douglas Adams: What Do You Do If You Are a Manically Depressed Robot?

- Adrian Currie appeared on the Naked Scientists radio show, on the episode Planet B: Should we leave Earth?

- Adrian also held a book launch for *Rock, Bone and Ruin: An Optimist's Guide to the Historical Sciences* at the Whipple Museum of the History of Science in May.

- Shahar Avin gave a podcast interview to Calcalist, Israel's most popular economic daily newspaper.

- Catherine Rhodes gave a 'Minerva Talk' on Science, Society and the End of the World, at St James Senior Girls School, London.

- Vision publisher David Hulme spoke to Seán Ó hÉigeartaigh about AI and existential risk, and released an article and hour-long video interview.

5. Recruitment and research team:

- We have just appointed a new **Research Project Administrator –** Clare Arnstein, who will start in early December, and is currently Executive Assistant to the Vice Chancellor (on secondment from the School of Arts and Humanities). We have also just recruited an additional Senior Research Associate as an **Academic Programme Manager.**

New Postdoctoral Research Associates:

- **Dr Luke Kemp** will work on the horizon-scanning and foresight strand of the Managing Extreme Technological Risks project. Luke has a background in international relations, particularly in relation to climate change policy and negotiations, and has been working recently as an economics consultant. Luke is interested in applying systems approaches to forecasting of extreme technological risks, and matching with mitigation and prevention strategies.

- **Dr Lauren Holt** will work on biological risks, in particular providing support for Lalitha Sundaram on the new Schmidt Sciences project on Extreme Risks from Chronic Disease Threats. Lauren has a background in zoology and applied ecology. Joins us from the Environment and Sustainability Institute at the University of Exeter. Lauren's also been involved with science communication and public engagement projects, and is planning to develop a career in science policy.

- **Asaf Tzachor** is expected to joining us for about a year. He will work on a project on food security, vulnerabilities in the global food system, and global catastrophic risk scenarios. He recently finished his doctorate at UCL's Department of Science, Technology, Engineering and Public Policy as a Goldman Scholar. He is a Fellow of the Royal Geographical Society (RGS), and was Head of Strategy and Sustainability at the Ministry of Environment (Israel). He has also written and edited a dozen national reports, books, academic articles, and government resolutions. He has also taught in the Interdisciplinary Center Herzliya, School of Sustainability and School of Government (Israel's top-ranked private college).

Visiting researchers:

- **Sam Weiss-Evans**, Assistant Professor in the Program on Science, Technology and Society at Tufts University, is visiting CSER from September 2018 – July 2019, and will be completing a book manuscript on the governance of security concerns in science and technology. Sam is also working to build a collaboration between CSER, MIT and the US National Academies on innovative approaches to governing dual use research.

New CSER Research Affiliates:

- **Adrian Currie**, left CSER in September for a lectureship in Philosophy at Exeter University, but continues to collaborate with CSER and CFI on science and creativity.

- **Daikichi Seki**, is a JSPS funded PhD student at the Graduate Institute of Advanced Integrated Studies in Human Survivability (GSAIS), Kyoto University. He is planning a visit to DAMPT next year to work on solar aspects of space weather, and will also spend some time with CSER to reflect on social aspects of the issue. This will be an initial phase in collaboration between GSAIS and CSER, with a plan to make a joint application to the Nippon / Sasakawa Foundation.

- **Yasmine Rix** has been actively engaged with CSER's work over the past few years, and will be curating an exhibition 'Ground Zero Earth' in the Alison Richard Building in February and March 2019, which will connect themes of CSER's research to the work of several emerging artists. She has secured in kind support from CRASSH, and funding from Cambridge Business Innovation District. She will help us run a public panel at the launch of the exhibition and will be doing some school engagement work as well.

- **Zoe Cremer** is a visiting student from ETH Zurich based at CFI for the 2018/2019 academic year, working with Sean O hEigeartaigh and Marta Halina on models of progress in artificial intelligence. Her work intersects with a number of CSER topics, and she will be a regular participant in CSER research meetings.

- **Tatsuya Amano** will be leaving in January to become a prestigious Australian Research Council Future Fellow at the University of Queensland. When he does so, we intend to propose him as a Research Affiliate.

6. Expert Workshops and Public Lectures:

Our events over the last few months have included:

- July: **Decision Theory & the Future of Artificial Intelligence** Workshop (led by Huw Price and Yang Liu). Held in Munich, it was the second in a workshop series that brings together philosophers, decision theorists, and AI researchers in order to promote decision theory research that could help make AI safer. It consolidated our partnership with the Munich Center for Mathematical Philosophy, a leader in this area.

- September: **Workshops with the Singaporean Government.** CSER, CFI and the Centre for Strategic Futures (part of the Singaporean Prime Minister's Office) co-organised a series of workshops in Singapore that explored existential risk, foresight, and AI. It helped consolidate our relationship with the Singaporean Government, an influential and far-sighted global player.

- September: **Plutonium, Silicon and Carbon** Workshop (led by Shahar Avin). It explored cybersecurity risks to nuclear weapons systems in the context of advances in AI and machine learning. It might lead to a paper with key experts from nuclear security, AI and cybersecurity. It also furthered collaboration with the United Nations Disarmament Research Centre - CSER researchers will visit UNIDIR in Geneva in November.
- Followed by a Public Lecture by **Dr Wade Huntley** on 'North Korea's Nuclear Policy'. Dr Wade Huntley teaches at the US Naval Postgraduate School and has published work on US strategic policies, East and South Asian regional security, and international relations theory.

- October: **Epistemic Security** Workshop (led by Shahar Avin). This began a series of workshops co-organised with the Alan Turing Institute, looking at the changing threat landscape of information campaigns and propaganda, given current and expected advances in machine learning.

- October: **Generality and Intelligence: from Biology to AI** Workshop (led by Seán Ó hÉigeartaigh). It explored how to evaluate progress in artificial intelligence in the context of different definitions of generality. It began the Cambridge² workshop series that will take place in Cambridge, UK, and Cambridge, MA, in the following two years, co-organised by the MIT-IBM Watson AI Lab and the Leverhulme Centre for the Future of Intelligence. MIT is a major player in AI research and development, recently launching a $1bn new school for AI.

- October: Public Lecture by **Dr Eli Fenichel** on 'Developments in the measurement of natural capital to advance sustainability assessment'. Dr Fenichel is an Associate Professor at Yale University. This lecture was co-organised with the Cambridge Conservation Initiative.

7. Upcoming activities

Four books will be published in early 2019:

- **Extremes**, 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.

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

- **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.

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

Upcoming events:

- November, January: **Epistemic Security** Workshop (led by Shahar Avin). Next in the series of workshops co-organised with the Alan Turing Institute, looking at the changing threat landscape of information campaigns and propaganda, given current and expected advances in machine learning.

- January: We are co-organising the **SafeAI 2019** Workshop, the Association for the Advancement of Artificial Intelligence's (AAAI) Workshop on AI Safety.

- February/March: **Ground Zero Earth Art Exhibition**. We are collaborating with Yasmine Rix on this art exhibition at the Alison Richard Building, to engage academics and the public in our research. The launch event will be on the evening of the 14 February.

Timing to be confirmed:

- Spring: **Cost-benefit Analysis of Technological Risk** Workshop (led by Simon Beard).

- Spring: **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 the Leverhulme Centre for the Future of Intelligence. MIT is a major player in AI, recently launching a $1bn new school for AI.

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

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

- Summer: **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 in order to promote research at the nexus between decision theory and AI. Co-organised with the Munich Center for Mathematical Philosophy.

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

- Public lectures: we will continue to hold at least six public lectures each year. Most of these will link to one of our workshops.

8. Publications

Adrian Currie (ed.) (2018) Special Issue: Futures of Research in Catastrophic and Existential Risk. *Futures.*

Many of the fifteen papers in the Special Issue were originally presented at our first Cambridge Conference on Catastrophic Risk in 2016, and it includes three papers by CSER researchers:

- **Adrian Currie, Seán Ó hÉigeartaigh**. (2018). Working together to face humanity's greatest threats: Introduction to the Future of Research on Catastrophic and Existential Risk. *Futures.*
- **Shahar Avin, Bonnie Wintle, Julius Weitzdörfer, Seán Ó hÉigeartaigh**, **William Sutherland, Martin Rees.** (2018). Classifying Global Catastrophic Risks. *Futures*.

> "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."

- **Adrian Currie**. (2018). Geoengineering Tensions. *Futures*.

> "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."

The rest of the papers are:

- Hin-Yan Liu, Kristian Cedervall Lauta, Matthijs Michiel Maas. (2018). Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures.*
- Alexey Turchin, David Denkenberger. (2018). Global catastrophic and existential risks communication scale. *Futures.*

- Peter Kareiva, Valerie Carranza. (2018). Existential risk due to ecosystem collapse: Nature strikes back. *Futures.*
- David Denkenberger, Robert W Blair Jr. (2018). Interventions that may prevent or mollify supervolcanic eruptions. *Futures.*
- John Halstead. (2018). Stratospheric aerosol injection research and existential risk. *Futures.*
- Donghyun Kim, Seul-Ki Song. (2018). Measuring changes in urban functional capacity for climate resilience: Perspectives from Korea. *Futures.*
- Jo L. Husbands. (2018). The challenge of framing for efforts to mitigate the risks of "dual use" research in the life sciences. *Futures.*
- Christine Aicardi, B. Tyr Fothergill, Stephen Rainey, Bernd Carsten Stahl, Emma Harris. (2018). Accompanying technology development in the Human Brain Project: From foresight to ethics management. *Futures.*
- Michael Crowley, Lijun Shang, Malcolm Dando. (2018). Preserving the norm against chemical weapons: A civil society initiative for the 2018 4th review conference of the chemical weapons convention. *Futures.*
- Denis Binder. (2018). The findings of an empirical study of the application of criminal law in non-terrorist disasters and tragedies. *Futures.*
- Claire Craig. (2018). Risk management in a policy environment: The particular challenges associated with extreme risks. *Futures.*
- Natalie Jones, Mark O'Brien, Thomas Ryan. (2018). Representation of future generations in United Kingdom policy-making.

Scientific communities and existential risk

- **Catherine Rhodes.** Scientific freedom and responsibility in a biosecurity context. Chapter 6 in Simona Giordano (Ed.) (2018). *The freedom of scientific research: Bridging the gap between science and society.* Manchester University Press.
  - "Scientific freedoms are exercised within the context of certain responsibilities, which in some cases justify constraints on those freedoms. (Constraints that may be internally established within scientific communities and/or externally enacted.) Biosecurity dimensions of work involving pathogens are one such case and raise complex challenges for science and policy. The central issues and debates are illustrated well in the development of responses to publication of ('gain of function') research involving highly pathogenic avian influenza, by a number of actors, including scientists, journal editors, scientific academies, and national and international policy groups."

- **Adrian Currie** (ed.). (2018). Special Issue Creativity, Conservatism & the Social Epistemology of Science. *Studies in the History and Philosophy of Science.*
  - **Adrian Currie.** (2018). Introduction. *Studies in the History and Philosophy of Science.*
    - "The special issue Creativity, Conservatism & the Social Epistemology of Science collects six papers which, in different ways, tackle 'promotion questions' concerning scientific communities: which features shape

those communities, and which might be changed to promote the kinds of epistemic features we desire. In this introduction, I connect these discussions with more traditional debate in the philosophy of science and reflect upon the notions of creativity which underwrite the papers."

  - **Adrian Currie**. (2018). Existential Risk, Creativity & Well-Adapted Science. *Studies in the History and Philosophy of Science.*
    - "Existential risks, particularly those arising from emerging technologies, are a complex, obstinate challenge for scientific study. This should motivate studying how the relevant scientific communities might be made more amenable to studying such risks. I offer an account of scientific creativity suitable for thinking about scientific communities, and provide reasons for thinking contemporary science doesn't incentivise creativity in this specified sense. I'll argue that a successful science of existential risk will be creative in my sense. So, if we want to make progress on those questions we should consider how to shift scientific incentives to encourage creativity. The analysis also has lessons for philosophical approaches to understanding the social structure of science. I introduce the notion of a 'well-adapted' science: one in which the incentive structure is tailored to the epistemic situation at hand."

Government reactions to disasters

- Elisa Hörhager, **Julius Weitzdörfer**. From Natural Hazard to Man-Made Disaster: The Protection of Disaster Victims in China and Japan. In Iwo Amelung et al. (Eds.) (2018). *Protecting the Weak in East Asia: Framing, Mobilisation and Institutionalisation.* Routledge.
  - "In East Asia, disasters have been regarded as events which uncover the mistakes of the past as much as they provide opportunities for building a more just society. In Japan, this phenomenon was captured through the concept of "world rectification" (yonaoshi) in the past and continues to lead to the improvement of disaster preparedness to this day. In the same way, disasters in historical China were not only interpreted as expressions of heavenly wrath for a ruler's mistakes, but also as an opportunity for better governance. Taking into account the way in which disasters simultaneously mirror existing trajectories and open up space for new ones, this chapter compares the protection of disaster victims in China and Japan by looking at two recent catastrophes, the 2008 earthquake in Wenchuan and the earthquake, tsunami and nuclear meltdown of 11 March 2011 in eastern Japan. We pay particular attention to the framing of both disasters as either man-made or natural, which carries significant social and political implications. Both governments made use of this distinction to shrug off responsibility and to influence mobilisation processes among the victims. The distinction between man-made and natural disasters also had a significant influence on the resulting institutionalisation processes."

Environmental assessment of high-yield farming

- Andrew Balmford, **Tatsuya Amano**, Harriet Bartlett, Dave Chadwick, Adrian Collins, David Edwards, Rob Field, Philip Garnsworthy, Rhys Green, Pete Smith, Helen Waters, Andrew Whitmore, Donald M. Broom, Julian Chara, Tom Finch, Emma Garnett, Alfred Gathorne-Hardy, Juan Hernandez-Medrano, Mario Herrero, Fangyuan Hua, Agnieszka Latawiec, Tom Misselbrook, Ben Phalan, Benno I. Simmons, Taro Takahashi, James Vause, Erasmus zu Ermgassen, Rowan Eisner. (2018). The environmental costs and benefits of high-yield farming. *Nature Sustainability.*
  - "How we manage farming and food systems to meet rising demand is pivotal to the future of biodiversity. Extensive field data suggest that impacts on wild populations would be greatly reduced through boosting yields on existing farmland so as to spare remaining natural habitats. High-yield farming raises other concerns because expressed per unit area it can generate high levels of externalities such as greenhouse gas emissions and nutrient losses. However, such metrics underestimate the overall impacts of lower-yield systems. Here we develop a framework that instead compares externality and land costs per unit production. We apply this framework to diverse data sets that describe the externalities of four major farm sectors and reveal that, rather than involving trade-offs, the externality and land costs of alternative production systems can covary positively: per unit production, land-efficient systems often produce lower externalities. For greenhouse gas emissions, these associations become more strongly positive once forgone sequestration is included. Our conclusions are limited: remarkably few studies report externalities alongside yields; many important externalities and farming systems are inadequately measured; and realizing the environmental benefits of high-yield systems typically requires additional measures to limit farmland expansion. Nevertheless, our results suggest that trade-offs among key cost metrics are not as ubiquitous as sometimes perceived."

Issues in decision theory relevant to advanced artificial intelligence:

- **Yang Liu, Huw Price**. (2018). Ramsey and Joyce on deliberation and prediction. *Synthese.*
  - "Can an agent deliberating about an action A hold a meaningful credence that she will do A? 'No', say some authors, for 'deliberation crowds out prediction' (DCOP). Others disagree, but we argue here that such disagreements are often terminological. We explain why DCOP holds in a Ramseyian operationalist model of credence, but show that it is trivial to extend this model so that DCOP fails. We then discuss a model due to Joyce, and show that Joyce's rejection of DCOP rests on terminological choices about terms such as 'intention', 'prediction', and 'belief'. Once these choices are in view, they reveal underlying agreement between Joyce and the DCOP-favouring tradition that descends from Ramsey. Joyce's Evidential Autonomy Thesis is effectively DCOP, in different terminological clothing. Both principles rest on the so-called 'transparency' of first-person present-tensed reflection on one's own mental states."

Theoretical mapping of artificial intelligence

- Fernando Martínez-Plumed, Bao Sheng Loe, Peter Flach, **Seán Ó hÉigeartaigh**, Karina Vold, José Hernández-Orallo. (2018). The Facets of Artificial Intelligence: A Framework to

Track the Evolution of AI. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).*

- "We present nine facets for the analysis of the past and future evolution of AI. Each facet has also a set of edges that can summarise different trends and contours in AI. With them, we first conduct a quantitative analysis using the information from two decades of AAAI/IJCAI conferences and around 50 years of documents from AI topics, an official database from the AAAI, illustrated by several plots. We then perform a qualitative analysis using the facets and edges, locating AI systems in the intelligence landscape and the discipline as a whole. This analytical framework provides a more structured and systematic way of looking at the shape and boundaries of AI."

- Fernando Martínez-Plumed, **Shahar Avin**, Miles Brundage, Allan Dafoe, **Seán Ó hÉigeartaigh**, José Hernández-Orallo. (2018). Accounting for the Neglected Dimensions of AI Progress. *arXiv.*
  - "We analyze and reframe AI progress. In addition to the prevailing metrics of performance, we highlight the usually neglected costs paid in the development and deployment of a system, including: data, expert knowledge, human oversight, software resources, computing cycles, hardware and network facilities, development time, etc. These costs are paid throughout the life cycle of an AI system, fall differentially on different individuals, and vary in magnitude depending on the replicability and generality of the AI solution. The multidimensional performance and cost space can be collapsed to a single utility metric for a user with transitive and complete preferences. Even absent a single utility function, AI advances can be generically assessed by whether they expand the Pareto (optimal) surface. We explore a subset of these neglected dimensions using the two case studies of Alpha* and ALE. This broadened conception of progress in AI should lead to novel ways of measuring success in AI, and can help set milestones for future progress."

- Sankalp Bhatnagar, Anna Alexandrova, **Shahar Avin**, Stephen Cave, Lucy Cheke, Matthew Crosby, Jan Feyereisl, Marta Halina, Bao Sheng Loe, **Seán Ó hÉigeartaigh**, Fernando Martínez-Plumed, **Huw Price**, Henry Shevlin, **Adrian Weller**, Alan Winfield, José Hernández-Orallo. (2018). Mapping Intelligence: Requirements and Possibilities. *In: Müller V. (eds) Philosophy and Theory of Artificial Intelligence 2017. PT-AI 2017. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol 44. Springer, Cham.*
  - "New types of artificial intelligence (AI), from cognitive assistants to social robots, are challenging meaningful comparison with other kinds of intelligence. How can such intelligent systems be catalogued, evaluated, and contrasted, with representations and projections that offer meaningful insights? To catalyse the research in AI and the future of cognition, we present the motivation, requirements and possibilities for an atlas of intelligence: an integrated framework and collaborative open repository for collecting and exhibiting information of all kinds of intelligence, including humans, non-human animals, AI systems, hybrids and collectives thereof. After presenting this initiative, we review related efforts and present the requirements of such a framework. We survey existing visualisations and representations, and discuss which criteria of inclusion should be used to configure an atlas of intelligence."