**Institution:** University of Cambridge

**Unit of Assessment:** UoA 30 Philosophy

**Title of case study:** Existential Risk: Building a Field and Influencing Policy

**Period when the underpinning research was undertaken:** 2012 – Present

**Details of staff conducting the underpinning research from the submitting unit:**

| Name(s): | Role(s) (e.g. job title): | Period(s) employed by submitting HEI: |
|---|---|---|
| Huw Price | Professor of Philosophy | October 2011 – Present |
| Seán Ó hÉigeartaigh | Programme Director, AI: Futures and Responsibility | Apr 2015 – Present |
| Shahar Avin | Senior Research Associate, CSER | Sep 2015 – Present |
| Catherine Rhodes | Senior Research Associate | Jan 2016 – Present |
| Partha Dasgupta | Professor Emeritus of Economics | 1992 – Present (Emeritus from 2010) |
| Adrian Currie | Research Associate | June 2016 – 31 August 2018 |
| Simon Beard | Senior Research Associate | April 2016 – Present |
| Natalie Jones | Research Associate | Oct 2019 – Present |
| Haydn Belfield | Research Associate | Jan 2017 – Present |

**Period when the claimed impact occurred:** 2014 – present

**Is this case study continued from a case study submitted in 2014?** No

## 1. Summary of the impact (indicative maximum 100 words)

The Centre for the Study of Existential Risk (CSER) is dedicated to the study and mitigation of risks that could lead to human extinction or civilisational collapse. Thanks to the Centre´s research and lobbying activity, governments, policymakers, and AI businesses around the world have increased their attention to, and introduced measures to reduce, existential risk. CSER researchers have helped to grow and shape the field by advising a range of new non-academic research centres and philanthropic funders on these emerging areas of risk research. The team has had a significant effect on UK and international policy by creating a new All-Party Parliamentary Group on Future Generations; by inspiring a campaign for a new UK Future Generations Bill; and by changing international norms regarding the publication of AI-technology research and the conduct of risk-assessments.

## 2. Underpinning research (indicative maximum 500 words)

CSER is an interdisciplinary research centre, founded in November 2012 by Huw Price, the Bertrand Russell Professor of Philosophy; Lord Martin Rees, Astronomer Royal; and Jaan Tallinn, co-founder of Skype. CSER research addresses theoretical, methodological and practical questions common to all potentially catastrophic risks [R1]. Its research also addresses specific questions concerning, for example, environmental risks [R2], and risks from AI [R5]. This work draws heavily on philosophical research, particularly in ethics, political philosophy, and the philosophy of science. All impact selected for inclusion in this case study was made possible thanks to the research conducted by CSER philosophers.

CSER’s work in ethics and political philosophy includes research on our obligations to future generations [R3]. A catastrophe on the scale of the Black Death would be a moral disaster for the current generation. However, if civilization were to be drastically derailed, then this would also have a negative effect on future generations. Future generations do not have a voice or a vote in current society. CSER researchers have argued this is an ethical problem that demands attention.

CSER's research in philosophy of science has generated insights that have been translated into recommendations for the emerging field of existential risk research, influencing new research centres and their approach to new science on risk. Currie, for instance, has argued that while existential risk research should incentivise a specific form of creativity, in fact most traditional sciences have incentive structures that are poorly adapted to this requirement [R4].

CSER has particular expertise in 'structured expert elicitation' methodologies; consulting experts where data and evidence are insufficient to forecast risks, but in a way that controls for the possibility of expert bias. This method has resulted in an influential report on 'The Malicious Use of Artificial Intelligence' [R5]. That report—the result of a workshop co-organised by CSER and the University of Oxford's Future of Humanity Institute—surveys and proposes ways to mitigate security threats from AI and machine learning. This method has also been deployed by Sir Partha Dasgupta, Chair of CSER, in the organisation of a major workshop on the ethics of climate change and resulting outputs with the Pontifical Academy of Sciences and the Pontifical Academy of Social Sciences at the Vatican [R2, R6].

## 3. References to the research (indicative maximum of six references)

[R1] Avin, S, Wintle, B, Weitzdörfer, J., O hÉigeartaigh, S, Sutherland, W. and Rees, M. (2018) 'Classifying Global Catastrophic Risk.' *Futures* 102: 20-26*.*

[R2] Dasgupta, P. and Ramanathan, V. (2014). 'Pursuit of the common good.' *Science* 345 (6203), 1457-1458.

[R3]. Beard, B., (2019) 'What is Unfair about Unequal Brute Luck? An Intergenerational Puzzle.' *Philosophia* 47(4).

[R4] Currie, A. (2018). 'Existential Risk, Creativity and a Well-Adapted Science*.' Studies in History and Philosophy of Science* 76: 39-48*.*

[R5] Brundage, M., Avin, S et al. (2018). 'The Malicious Use of Artificial Intelligence: Forecasting, Preventing and Mitigation'. *arXiv*.

[R6] Dasgupta, P., Ramanathan, V. and Sánchez Sorondo, M., eds. (2015) *Sustainable Humanity, Sustainable Nature: Our Responsibility*. Pontifical Academy of Sciences.

The above outputs have all been peer reviewed and thus meet the 2* requirement.

## 4. Details of the impact (indicative maximum 750 words)

**Building the field of existential risk**

CSER has been at the forefront of a developing area of academic and non-academic research into existential risk, influencing several new non-academic research centres, think tanks, and funding initiatives. It has organised two international conferences and over 35 expert workshops bringing together CSER researchers and other academics with representatives from national governments, international governing bodies, third sector organisations, think tanks, and technologists. CSER has also hosted over 60 visitors. Thanks to the profile of CSER's work among this emerging research and policy community, the team have been consulted by many new non-academic research centres as they found and establish the remit for their own work on existential risk. CSER has influenced:

- The Future of Life Institute: a volunteer-run research and outreach organisation based in Boston, USA, working to mitigate existential risk [E1]
- The Nuclear Threat Initiative (NTI): CSER researchers contributed to a collaboration between the St Catharine's College Biorisk Initiative (BioRISC) and NTI, which Lord Browne (former UK Defence Secretary and NTI Vice-Chair) says has 'significantly

increased their respective influences on biosecurity policy leadership in the Euro-Atlantic space' [E2]
- The Global Challenges Foundation (GCF): in 2019, CSER was commissioned by GCF to write two reports charting the current international governance of global catastrophic risk and its main drivers, contributing to GCF's policy research and its recommendations to improve global risk governance [E3]

**UK Future Generations APPG and Bill**

One of the most tangible consequences of CSER policy engagement has been the creation of a new UK All-Party Parliamentary Group (APPG) for Future Generations, and the drafting of a new Future Generations Bill. A 2018 paper in a *Futures* special issue (edited by Currie), co-authored by Natalie Jones (then a Cambridge PhD student and CSER research affiliate), and supervised by CSER researchers, specifically recommended establishing an APPG for Future Generations. On the basis of this recommendation, CSER researchers created that APPG in collaboration with MPs and peers from across the political spectrum. CSER hosted the APPG's secretariat in its first year. Since 2019 Belfield has acted as the liaison between CSER and the APPG [E4]. In the words of the coordinator and secretariat member of the APPG: 'Without CSER's research on representing future generations, the APPG for Future Generations would not exist' [E4]. The APPG has held five events in Parliament on the theme 'Managing Technological Risks', which have featured talks from CSER researchers Shahar Avin, Simon Beard, Catherine Rhodes, and Julius Weitzdörfer. Simon Beard also wrote the APPG report summarising findings from the event series [E4].

The APPG's activity is already having an effect in UK parliament and policymaking circles. In late 2019, the APPG (in collaboration with CSER) launched an Inquiry into 'Long-termism in Policy-making'. The APPG has informed key individuals in the UK Government (e.g. advising civil servants at the Department for Digital, Culture, Media and Sport) and it has been mentioned in Parliamentary debates. CSER and the APPG continue to work collaboratively, with Haydn Belfield as primary liaison [E4]. Together they have advised on the creation and direction of the Today For Tomorrow cross-party campaign. This campaign includes a draft Future Generations Bill, led by Lord John Bird. Lord Bird joined the APPG as co-Chair in Feb 2019, and visited the CSER team in Cambridge on 1 Mar 2019 to discuss the launch of the campaign and the Bill. CSER researchers advised on the drafting on the Bill, and some of the recommendations of the CSER paper in *Futures* have been included in it. Lord Bird has confirmed that without the Centre's work his campaign would not have started, and draft bill would not exist [E5].

**Putting existential risk on the global policy agenda**

The CSER team also regularly organises, and is invited to, conferences, workshops, and ad hoc meetings with UK and international policymakers. CSER researchers have thereby put the topic of existential risk on the agenda of high-level policy meetings, and have made specific recommendations for policy across a range of risks. Examples of this form of policy impact include:

- A significant influence on the European Commission's *White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, specifically via two written submissions, alongside extensive discussions with the Commission's High-Level Expert Group [E6]
- Attendance at invitation-only policy stakeholder events: e.g., a policy conference on Modern Deterrence at Ditchley Park in November 2018 attended by Shahar Avin, and a policy meeting in November 2018 on high impact bio-threats held at Wilton Park [E7]
- Contributions to UN negotiations on Lethal Autonomous Weapons Systems (2018), and to the Biological Weapons Convention annual meeting of states parties (2017 and 2018) [E8]
- Inspiring the Finnish government to commission the report *Existential Risk: Diplomacy and Governance* [E9]

One particularly notable example of CSER's engagement in policy issues is Sir Partha Dasgupta's work at the Vatican. In May 2014, Sir Partha co-organised a major workshop with the Pontifical Academy of Sciences. After the workshop, Sir Partha spoke to the Pope directly and encouraged him to include climate change in his speeches and to urge people to be better stewards of the planet. The workshop underpinned a major report [R6] published in April 2015 by the Vatican, which listed Dasgupta as one of its four corresponding authors in recognition of his contributions. The report in turn partly informed the May 2015 *Laudato si'* Papal Encyclical, which focussed on the impending threat of climate change [E10]. Sir Partha is now leading the UK Government's independent 'Dasgupta Review: The Economics of Biodiversity'.

**AI risks: forecasting, prevention, and mitigation**

CSER's research output *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* [R5], is the first major report to examine emerging risks at the intersection of artificial intelligence, cybersecurity, physical security and information manipulation. It is a collaborative report co-written with the Future of Humanity Institute, Oxford. It has been cited in reports from the UK House of Lords and the US Government Departments, and has received praise from policy-makers, technologists, a UK Minister, the Commander of the Australian Defence College, and the former President of the Association for the Advancement of Artificial Intelligence (AAAI) [E11].

This report has influenced policy discussions around AI and security. One example of this was a workshop series (prompted by the report) on 'Epistemic Security', investigating the changing threat landscape of information campaigns and propaganda given current advances in machine learning. These workshops were co-organised by CSER, the Alan Turing Institute, and the Defence Science and Technology Laboratory (Dstl, the UK government's leading experts in technology and security).

*Malicious Use* is changing the behaviour of non-academic AI research organisations. It encourages greater care and patience in the publication of results, and suggests that pre-publication risk assessments may be necessary in some circumstances. Since the report's publication, OpenAI (a major AI research company) has started to implement this new norm. In early 2019, it delayed publication of some of its results in accordance with the CSER recommendation. This decision started a widespread debate – with consequent behaviour change – across the AI community [E12]. Members of the CSER team subsequently worked with Partnership on AI – the leading non-profit coalition of AI technologists – to develop a more detailed report on AI Publication Norms for use by AI research companies [E12].

**5. Sources to corroborate the impact** (indicative maximum of 10 references)

[E1] Testimonial from co-Founder and President of the Future of Life Institute

[E2] Testimonial from Vice-chair, Nuclear Threat Institute.

[E3] The Cartography of Global Catastrophic Risk Governance, Assessing the Drivers of Global Catastrophic Risk.

[E4] Testimonial statement from coordinator and secretariat member for the APPG on Future Generations

[E5] Testimonial from Co-chair, All-party Parliamentary Group for Future Generations

[E6] Copy of the advice submitted to the Commission

[E7] Copies of webpages with details of Ditchley Park and Wilton Park events

[E8] Final report of UN Meeting

[E9] Testimonial from Professor of International Relations, Tampere University

[E10] *Laudato si'* On Care For Our Common home

[E11] The Malicious AI Report is cited in: the House of Lords Select Committee on Artificial Intelligence Report of Session 2017-19 (fn.9); 'AI: Using standards to mitigate risks', report by Public-Private Analytic Exchange Program, a partnership led by the US Department of Homeland Security and Defense Intelligence Agency (fn.3). PDFs provided.

[E12] Testimonial from Program Lead, Partnership on AI.