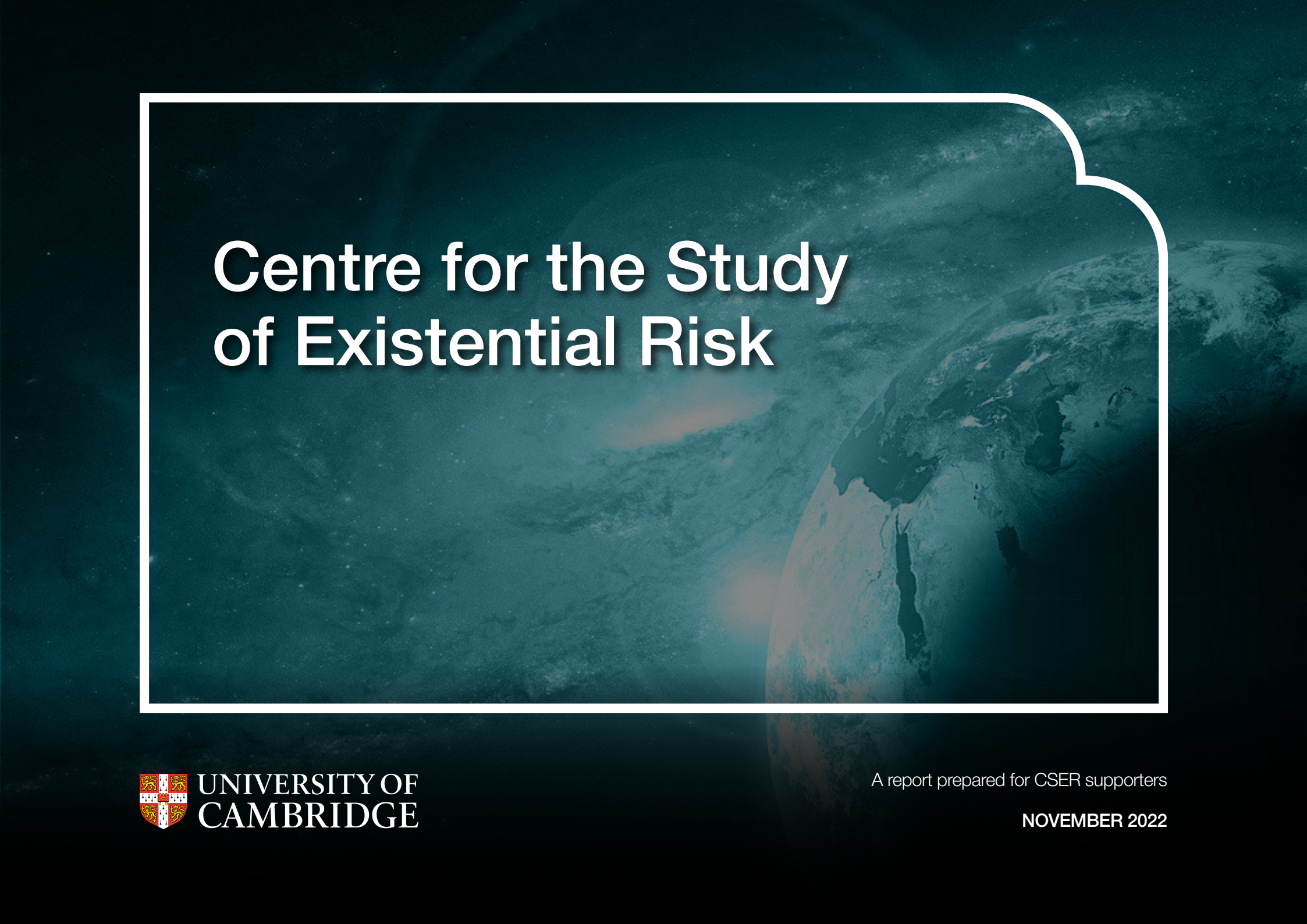


Centre for the Study of Existential Risk

UNIVERSITY OF CAMBRIDGE

A report prepared for CSER supporters

NOVEMBER 2022

The University of Cambridge extends its sincere thanks for your support of the activities of the Centre for the Study of Existential Risk (CSER).

Supported by your generosity, the work of CSER researchers is increasing our understanding of, and preparedness for, existential threats to our world.

Navigation

Scroll through the document, or click on the relevant section in the table of contents to go directly to that section. To return to the contents list, click the page number at the bottom of the page.

Click to return to contents

Centre for the Study of Existential Risk Q2 2022

Contents

An introduction from **Seán Ó hÉigeartaigh**

Executive Director, CSER

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 work primarily on catastrophic biological risks, environmental risks, and on risks from artificial intelligence, as well as on cross-cutting methodologies for the analysis and governance of global risks. Our work is shaped around three main goals:

- **Understanding**: we study existential and global catastrophic risk.
- **Impact**: we develop collaborative strategies to reduce these risks.
- **Field-building**: we foster a global community of academics, technologists and policy-makers who share our goals.

This report covers the period April–August 2022 and outlines our activities and future plans. Highlights of the last three months include:

- We published seven papers, two pre-prints, and a comment piece – including in ***Nature***, ***Nature Sustainability*** **and the** ***Proceedings of the National Academies of Science.*** These covered cooperation between Machine Learning agents, the political implications of developments in the life sciences, the safety of clinical applications of Machine Learning, AI performance metrics, the research agenda of the IPCC, catastrophic climate change, digital twins for sustainability, military applications of AI, the science of existential risk, and huge volcanic eruptions.
- We hosted a **highly successful conference**, a two-week programme of intensive engagement with a group of interdisciplinary international visitors, and an in-person public lecture by Professor Matthew Adler.
- Our researchers presented their work at 13 workshops, conferences, and talks hosted by, among others, the **European Geoscience Union**, the Vienna Complexity Science Hub, Darwin College, the Open University, the European Evaluation Society, the **Royal College of Defence Studies**, Walton Park, and **AI for Good**.
- Our work has also been featured in leading global media outlets, including ***The Independent***, BBC Radio 4, **Sky News**, Associated Press, TRT World News, *The Progressive*, and **BBC Future**. There was particular media interest in our researchers work on volcanic risk, catastrophic climate change, and the war in Ukraine.
- We welcomed four visitors working on AI safety, humanitarianism, and catastrophic risk and two of our researchers took up new positions with the **International Institute for Sustainable Development** and the **Legal Priorities Project**.

1. Update

CSER is now largely back to in person working and this period has seen us restart large scale in person events, including our fourth conference, a two week intensive visitor programme, and the first of our in-person public lectures since February 2020 by Professor Matthew Adler. We remain sensitive to individual risks and needs and open to the advantages of hybrid and remote working to boost productivity and wellbeing and continue to offer remote access to our events. We continue to plan towards our next stage of growth and the opportunities for new projects and research that more in-person working is making possible again and the process of recruiting a new Centre Director is now underway with an extended application deadline of 16 October.

2. People

2.1 Recruitment

CSER is currently recruiting for a Director (closing date for applications 16 October), Post-Doctoral Research Associate in AI risk and foresight (closing date for applications 25 September), and Research Assistant in Global Catastrophic Risk and Communication (closing date for applications 9 October).

2.2 Visiting Scholars

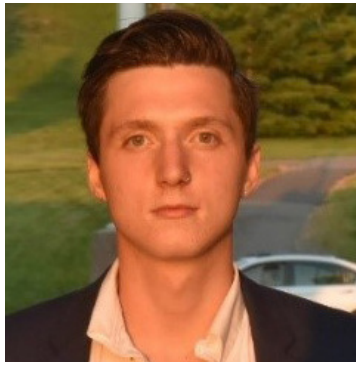
We have welcomed four visitors to CSER during this period.

Sumaya Nur is a final-year law student at Strathmore University in Nairobi.

Her research interests lie in the intersection of law and artificial intelligence. Sumaya's previous work has concentrated on the regulation of facial recognition algorithms, as well as artificial intelligence and liability. Her senior thesis will focus on broadening the causation test in establishing liability for decisions made by self-learning algorithms in common law jurisdictions. Sumaya intends to do her further studies in AI Safety and governance. She started her remote visit in June 2022.

Pablo Suarez is a system dynamics modeler turned humanitarian worker, innovator, game designer, and creator of serious-yet-fun processes for collaborative processes to inspire thinking and action. He is innovation lead at the Red Cross Red Crescent Climate Centre and artist in residence at the National University of Singapore. Pablo holds a water engineering degree, a master's in planning, and a PhD in geography. He will be visiting from July to November 2022.

Gordon Woo is a catastrophist, with research interests in all manner of catastrophes. He is the author of the books 'The Mathematics of Natural Catastrophes' and 'Calculating Catastrophe', published by World Scientific Press. For the past decade, he has focused on searching for ways of tracking surprising Black Swan events, and this has led to the development agenda of Downward Counterfactuals. He is a visiting professor at UCL, and an adjunct professor at NTU, Singapore. He visited in July 2022.



Coleman Snell is completing his undergraduate degree at Cornell University in Applied Ethics and Behavioral Psychology. Coleman runs the 21st Talks Podcast and Youtube channel where he interviews experts from x-risk and other related fields. Coleman also works as a researcher and office manager at Cornell's Long term Artificial Intelligence Safety Lab (LAISR), where he conducts research into behavioral biology applied to AI training, along with serving as President of Cornell Effective Altruism. Coleman's area of focus is in effective existential risk communication and motivation, through applying ideas from behavioral psychology and existential psychology to the problem of sustainable motivation to do the good. He visited in August 2022.

2.2 Research Affiliates

We have welcomed two new research affiliates:

Aaron Tang examines the feasibility, desirability, and governance of last-ditch efforts to address climate change (developing new technologies to pull greenhouse gases out of the atmosphere, and reflecting sunlight to cool the Earth). He is a PhD Scholar and Lecturer in Climate Policy at the Fenner School of Environment & Society at the Australian National University.

Ross Gruetzemacher works on forecasting and foresight methods related to technological progress and existential risk. His research focus is on the development and impacts of transformative AI.

2.3 Leavers

We have sadly said goodbye to two of our Research Associates. Natalie Jones, who has taken up a position as Policy Advisor on Sustainable Energy Supply at the International Institute for Sustainable Development, and Matthijs Maas, who has taken up a position as Senior Research Fellow (Law & AI) and Head of AI Research at the Legal Priorities Project. They will continue to collaborate with CSER as Research Affiliates, and we wish them all the best in their new roles.

3. Events, Engagement and Outreach

3.1 Academic engagement

Lara Mani co-authored a **highly prominent comment piece in *Nature*** calling for more research and greater preparedness for large volcanic eruptions. CSER researchers have presented research at workshops, conferences, and talks hosted by the European Geoscience Union, the Vienna Complexity Science Hub, Darwin College, the Open University, the European Evaluation Society, the Royal College of Defence Studies, Walton Park, and AI for Good among others, as part of our ongoing efforts to engage a wide range of academics in the need to understand and mitigate the biggest risks facing humanity. Sabin Roman presented his work on modelling societal evolution and collapse to a number of audiences as he prepares for significant publications in this space. Alongside her work promoting research into large scale volcanic eruptions, Lara Mani also gave several presentations of her work evaluating communications of volcanic risk, which feeds into her work on Global Catastrophic Risk (GCR) communication and outreach, and on evaluating existential risks. Other researchers presented on learning from the global south, paradigms for thinking about existential risk, disability and inclusion in Existential Risk Studies, trust and AI developers, and synthetic biology. CSER researchers were also acknowledged for their world building in **an international competition organised by the Future of Life Institute**.

- 6 May: Sabin Roman gave a talk on Modelling the long-term evolution of societies at the Complexity Science Hub in Vienna.
- 23 May: Lara Mani gave a talk on 'Evaluating the crisis communications campaign during the 2020–2021 eruption of La Soufriere, St Vincent' at EGU 2022.
- 24 May: Sabin Roman gave a talk on The Collapse of Complex Societies at Darwin College.
- 24 May: Lara Mani and Asaf Tzachor gave a talk on 'Lower magnitude volcanic eruptions as Global Catastrophic Risks' at the **European Geoscience Union's** 2022 conference (EGU 2022).
- 2 June: Sabin Roman gave a talk at the Dutch Institute for Emergent Phenomena on societal collapse.
- 7 June: SJ Beard and Paul Ingram presented on How to Think About the End of the World… and Approaches to Mitigate the Risks to the Open University Philosophy Faculty.
- 9 June: Lara Mani and Rick Davies gave a talk on Exploring And Evaluating Existential Risks: Process And Outcomes Of An International Participatory Exercise Examining Alternative Biotechnology Futures at the European Evaluation Society's 2022 conference (EES 2022).
- 14 June: Lara Mani gave a talk on 'Lessons from the 2020-2021 eruption of La Soufriere St Vincent on 'what works' for volcanic crisis communications" at the Cities on Volcanoes conference.
- 23 June: Clarissa Rios Rojas gave a talk on Latin-American scientists influencing global policy: lessons learnt from the Global South at the Sustainability Research and Innovation Congress 2022.
- 27 June: Paul Ingram and Shahar Avin gave talks about existential risks and nuclear threats at the **Royal College of Defence Studies**.

- 28 June: Seán Ó hÉigeartaigh spoke on the keynote panel on Existential Risk at the Global Priorities Institute's Annual Conference.
- 30 June: a team from CSER and the Leverhulme Centre for the Future of Intelligence were awarded second place in the **Future of Life Institute's** World Building Competition.
- 11 July: SJ Beard gave a talk on Disability and Existential Risk at the UK Transgender Philosophers summer school in Edinburgh.
- 14 July: Shahar Avin gave a talk on When should you trust the developers of AI systems? at the **AI for Good Neural Network.**
- 22 July: Lalitha Sundaram gave a talk on 'Safe, Secure, & Responsible Synthetic Biology Beyond Containment' at Walton Park.
- 17 August: Lara Mani co-authored a comment peace in **Nature** Huge volcanic eruptions: time to prepare with Mike Cassidy.
- Lalitha Sundaram mentored two students of existential risk, Hanna Pálya and Oscar Delaney, through the Cambridge Existential Risk Initiative Summer Fellowship scheme.

3.2 Policy Engagement

Our researchers continue to engage with policymakers at a variety of levels, including the UK's Foreign Affairs Select Committee, UN Global Platform for Disaster Risk Reduction, and UK Health Security Agency and Foreign Office. Clarissa Rios Rojas made particular inroads in engaging with the **UN office for Disaster Risk Reduction**, which also benefited from the presentation by Jenty Kirsch-Wood at our conference.

- 21 March: Seán Ó hÉigeartaigh met with then-Minister for Digital, Culture, Media and Sport Chris Philp and senior members of the Office for AI to advise on the UK's National AI strategy and academic collaboration.
- 12 April: Tom Hobson submitted written evidence to the UK Parliament's Foreign Affairs Select Committee Inquiry on Tech and the future of UK Foreign Policy.
- 25 May: Clarissa Rios Rojas gave a talk on Building a science-policy interface to tackle global governance of catastrophic & existential risk at the **UN Global Platform for Disaster Risk Reduction.**
- 26 May: Clarissa Rios Rojas provided a statement on behalf of CSER at the Midterm Review Plenary 2: Beyond natural hazards – operationalising the expanded scope of the Sendai Framework during the UN Global Platform for Disaster Risk Reduction.
- 10 June: Freya Jephcott met with CSaP fellow Scott McPherson, Director General, Strategy, Policy and Programmes, UK Health Security Agency.
- 23 June: CLTR researcher and CSER affiliate Jess Whittlestone, Shahar Avin, Di Cooke, Kayla Lucero-Matteucci, Seán Ó hÉigeartaigh, Haydn Belfield and the Centre for Governance of AI's Markus Anderljung published a response to the UK's Defence AI strategy.
- 20-22 June: Paul Ingram attended the 4th International Conference on the Humanitarian Consequences of Nuclear Weapons, followed by the First Meeting of States Parties to the Treaty on the Prohibition of Nuclear Weapons, both in Vienna.
- July: Paul Ingram was involved in meetings with officials from the White House, and then the Cabinet Office, alongside collaborators and scientists to discuss policy consequences arising from climatic effects of nuclear weapons.
- 8 August: CLTR researcher and CSER affiliate Jess Whittlestone, Shahar Avin and colleagues submitted evidence for the UK Government's Future of Compute Review.
- 1–12 August: Paul Ingram attended the Tenth NPT Review Conference in New York. This included giving a talk alongside the Chinese delegation leader and other diplomats, to discuss the Stepping Stones Approach to Nuclear Disarmament and the Stockholm Initiative.
- 12 August: CSER researchers met with Vijay Rangarajan, the **UK Foreign Office's Director General**: America, Afghanistan, Pakistan, Middle East, Overseas Territories and India.

3.3 Public Engagement

Our work has been showcased by leading global media outlets, including *The Independent*, BBC Radio 4, Sky News, Associated Press, TRT World News, *The Progressive*, and BBC Future, while our researchers have also engaged with a range of public outreach talks and events. In particular there was **significant media attention** around Lara Mani's work on volcanic risks, Luke Kemp's work on catastrophic climate change, and Paul Ingram's ongoing analysis of the war in Ukraine and the risk that this could lead to the use of nuclear weapons.

- 5 April: Clarissa Rios Rojas spoke on a panel on 'Women in Global Governance' at the Bucerias Summer School.
- 12 April: Matthijs Maas wrote a blog post for the Effective Altruism Forum: A primer & some reflections on recent CSER work.
- 27 April: Paul Ingram spoke about The Risk of Nuclear War to TRT World News.
- 26 May: Paul Ingram published an article Russia's Nuclear Threat Endangers Us All in *The Progressive* magazine.
- 29 May: Ellen Quigley appeared on Episode 46 of the Boglerheads On Investing podcast on ESG Investing, hosted by Rick Ferri.
- 9 June: Lauren Holt published an article Should we detach ourselves from nature? on **BBC Future**.
- 29 June: Paul Ingram appeared on an episode of BBC Radio 4's The Moral Maze on Ukraine – what should western countries do next?
- 2 July: Paul Ingram spoke about the NATO Summit, Russia and the threat of nuclear war on BBC Radio Scotland.
- 8 July: Lara Mani gave a talk as part of the EA communications fellowship programme.
- 14 July: Haydn Belfield published a blog post on Why policy makers should beware claims of new 'arms races' in the *Bulletin of Atomic Scientists*
- 2 August: Luke Kemp's paper on catastrophic climate change in *PNAS* received media attention from, amongst others, AP News: Chances of climate catastrophe are ignored, scientists say, **Sky News**: Catastrophic effects of climate change are 'dangerously unexplored', experts warn, and Axios: Climate change catastrophes need greater study, scientists warn.
- 3 August: SJ Beard and Paul Ingram appeared on an episode of BBC Radio 4's Sideways about the risk of nuclear war.
- 9 August: SJ Beard appeared on an episode of the Dommer Optimism podcast to talk about existential risk.
- 15 August: Paul Ingram was quoted by ***The Independent*** in an article Imagine there's been a nuclear attack. Here's how Britain should respond.
- 17 August: Lara Mani's *Nature Comment* piece received media attention from, amongst others, University of Cambridge: Risk of volcano catastrophe 'a roll of the dice', say experts and ***The Times***: One in six chance of a massive volcanic eruption this century.
- 18 August: Haydn Belfield and Shin-Shin Hua published a blog post Effective Enforceability of EU Competition Law Under Different AI Development Scenarios for *Verfassungsblog*.
- 21–29 August: Paul Ingram appeared on Al Jazeera, Talk TV, GBTV, Times Radio and TRT talking about the dangers at the Zaporizhzhia Nuclear Power Plant. See, for example, the TRT World News programme IAEA mission departs for Zaporizhzhia nuclear power plant

3.4 Events

We have been delighted to relaunch our in-person public events series this term, including a highly **successful hybrid format** conference, an intensive visitor programme, and a workshop and public lecture by Professor Matthew Adler.

- 19–21 April: CSER held our fourth Cambridge Conference on Catastrophic Risk at the Intellectual Forum, Jesus College, with three days of speakers, panels, workshops, and vibrant discussion. A summary of the event is included as an Appendix to this report.
- 22 April–5 May: CSER hosted 11 international visitors for a programme of **intensive interdisciplinary engagement** around global risks including a foresight exercise about the future of the field of Existential Risk Studies using the ParEvo technique.
- 23 May: Professor Matthew Adler (Richard A Horvitz Professor of Law and Professor of Economics, Philosophy and Public Policy at Duke University) gave a workshop on the philosophical foundations of fatality risk regulation followed by a public lecture on Measuring Social Welfare, with Priority for the Worst Off.

4. Publications

4.1 Papers

John Burden published a paper in the *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (SafeAI 2022)* conference about cooperation between Machine Learning agents:

Oases of Cooperation: An Empirical Evaluation of Reinforcement Learning in the Iterated Prisoner's Dilemma in *SafeAI@AAAI 2022* 13 March 2022 by Peter Barnett and John Burden.

> In the creation of safe AI systems it is extremely important to ensure cooperative behaviour of these systems, even when there are incentives to act selfishly. In many cases, even when game-theoretic solutions allow for cooperation, actually getting the AI systems to converge on these solutions through training is difficult. In this paper we empirically evaluate how reinforcement learning agents can be encouraged to cooperate (without opening themselves up to exploitation) by selecting appropriate hyperparameters and environmental perceptions for the agent. Our results in the multi-agent scenario indicate that in hyperparameter-space there are isolated "oases" of mutual cooperation, and small changes in these hyperparameters can lead to sharp drops into non-cooperative behaviour.

Tom Hobson co-authored a book chapter examining the philosophical and political implications or recent developments in the life sciences for our understanding of what it means to be human.

Questioning the Politics of Human Enhancement Technologies in Bioethics and the Posthumanities in *Bioethics and the Posthumanities* edited by Danielle Sands (Routledge) by Tom Hobson and Anna Roessing.

> Research in the biological sciences has revolutionised our understandings of life, biological entities, and boundaries of the organic and inorganic world, to the extent that extant ontologies of the human have been scrutinised. The discovery of the world of the microbiome substantiates calls for a post-anthropocentric turn in our definition of the human, her agency, and subjectivity. At the same time, the ability to change the genetic constitution of humans poses questions about the normative, ontological, and epistemological frameworks concerning the human essence and boundaries of normality and (dis)-ability within classical medical and philosophical explanatory models. From the desire to overcome death, to transcend the constraints of the human being and body – extending them to new human-machine interfaces – technological developments are situated in these new understandings but also actively shape visions of the human future. What remains underexplored within the field of political sciences is the role of technology in negotiating these imaginaries of societal futures. Technologies remain most frequently understood as either instrumental and ambivalent in regard to societal norms or deterministic in bringing dystopian or utopian ends to society. In this chapter, we look to reveal the commitments, values, and

norms embedded in the application of scientific knowledge and its materialisation in technological artefacts and practices in the field of human augmentation technologies. We scrutinise the stakes of contemporary imaginations of technological innovation as a force of social and historical transformation, social progress, and in defining the essence of human life. Attending to post-human and trans-human visions, we explore their utopian and eschatological dimensions and metaphors while placing post-human visions within a broader context of ideas, institutions, and practices of innovation.

John Burden co-authored a paper on the safety of clinical applications of Machine Learning in the *Journal of Biomedicine*.

Safety-driven design of machine learning for sepsis treatment in *Journal of Biomedical Informatics* by Yan Jia, Tom Lawton, John Burden, John McDermid, and Ibrahim Hablia.

Machine learning (ML) has the potential to bring significant clinical benefits. However, there are patient safety challenges in introducing ML in complex healthcare settings and in assuring the technology to the satisfaction of the different regulators. The work presented in this paper tackles the urgent problem of proactively assuring ML in its clinical context as a step towards enabling the safe introduction of ML into clinical practice. In particular, the paper considers the use of deep Reinforcement Learning, a type of ML, for sepsis treatment. The methodology starts with the modelling of a clinical workflow that integrates the ML model for sepsis treatment recommendations. Then safety analysis is carried out based on the clinical workflow, identifying hazards and safety requirements for the ML model. In this paper the design of the ML model is enhanced to satisfy the safety requirements for mitigating a major clinical hazard: sudden change of vasopressor dose. A rigorous evaluation is conducted to show how these requirements are met. A safety case is presented, providing a basis for regulators to make a judgement on the acceptability of introducing the ML model into sepsis treatment in a healthcare setting. The overall argument is broad in considering the wider patient safety considerations, but the detailed rationale and supporting evidence presented relate to this specific hazard. Whilst there are no agreed regulatory approaches to introducing ML into healthcare, the work presented in this paper has shown a possible direction for overcoming this barrier and exploit the benefits of ML without compromising safety.

John Burden and Jose Hernandez-Orallo co-authored a paper with colleagues from the Leverhulme Centre for the Future of Intelligence on alternatives to average performance metrics for AI systems that look at an agent's pattern of performance across different tasks instead.

Not a Number: Identifying Instance Features for Capability-Oriented Evaluation May 2022 by Ryan Burnel, John Burden, Danaja Rutar, Konstantinos Voudouris, Lucy Cheke, and Jose Hernandez-Orallo.

In AI evaluation, performance is often calculated by averaging across various instances. But to fully understand the capabilities of an AI system, we need to understand the factors that cause its pattern of success and failure. In this paper, we present a new methodology to identify and build informative instance features that can provide explanatory and predictive power to analyse the behaviour of AI systems more robustly. The methodology builds on these relevant features that should relate monotonically with success, and represents patterns of performance in a new type of plots known as 'agent characteristic grids'. We illustrate this methodology with the Animal-AI competition as a representative example of how we can revisit existing competitions and benchmarks in AI—even when evaluation data is sparse. Agents with the same average performance can show very different patterns of performance at the instance level. With this methodology, these patterns can be visualised, explained and predicted, progressing towards a capability-oriented evaluation rather than relying on a less informative average performance score.

Luke Kemp co-authored a paper examining how the climate change scenarios most often explored in research cited by the IPCC compare with their likelihood and find that researchers consistently focus on less likely safer scenarios instead of the most realistic and most dangerous scenarios.

Focus of the IPCC Assessment Reports Has Shifted to Lower Temperatures in *Earth's Future* 10(5) May 2022 by Florian U Jehn, Luke Kemp, Ekaterina Ilin, Christoph Funk, Jason R Wang, and Lutz Breuer.

> We focus on how different global temperature increases represented in IPCC reports have shifted over time. While the first four assessment reports had a roughly equal focus on temperatures above and below 2°C, the more recent fifth and sixth assessment reports have a considerably stronger focus on warming below 2°C. This is concerning as warming above 2°C is more likely given current emissions trajectories and is more influential on climate risk assessments.

Luke Kemp was lead author of a prestigious paper making the case for greater research into the most catastrophic potentials for climate change and setting out how this could be done.

Climate Endgame: Exploring catastrophic climate change scenarios in the *Proceedings of the National Academies of Science* by Luke Kemp, Chi Xu, Joanna Depledge, Kristie L Ebi, Goodwin Gibbins, Timothy A Kohler, Johan Rockström, Marten Scheffer, Hans Joachim Schellnhuber, Will Steffen, Timothy M Lenton 2 August 2022.

> Prudent risk management requires consideration of bad-to-worst-case scenarios. Yet, for climate change, such potential futures are poorly understood. Could anthropogenic climate change result in worldwide societal collapse or even eventual human extinction? At present, this is a dangerously underexplored topic. Yet there are ample reasons to suspect that climate change could result in a global catastrophe. Analying the mechanisms for these extreme consequences could help galvanize action, improve resilience, and inform policy, including emergency responses. We outline current knowledge about the likelihood of extreme climate change, discuss why understanding bad-to-worst cases is vital, articulate reasons for concern about catastrophic outcomes, define key terms, and put forward a research agenda. The proposed agenda covers four main questions: 1) What is the potential for climate change to drive mass extinction events? 2) What are the mechanisms that could result in human mass mortality and morbidity? 3) What are human societies' vulnerabilities to climate-triggered risk cascades, such as from conflict, political instability, and systemic financial risk? 4) How can these multiple strands of evidence – together with other global dangers – be usefully synthesized into an "integrated catastrophe assessment"? It is time for the scientific community to grapple with the challenge of better understanding catastrophic climate change.

Asaf Tzachor and Catherine Richards were co-authors of a paper exploring the risks and limitations of using digital simulations to help achieve sustainability and ways to overcome these in *Nature Sustainability*.

Potential and limitations of digital twins to achieve the Sustainable Development Goals in *Nature Sustainability* by Asaf Tzachor, Soheil Sabri, Catherine Richards, Abbas Rajabifard, Michele Acuto 5 August 2022.

> Could computer simulation models drive our ambitions to sustainability in urban and non-urban environments? Digital twins, defined here as real-time, virtual replicas of physical and biological entities, may do just that. However, despite their touted potential, digital twins have not been examined critically in urban sustainability paradigms – not least in the Sustainable Development Goals framework. Accordingly, in this Perspective, we examine their benefits in promoting the Sustainable Development Goals. Then, we discuss critical limitations when modelling socio-technical and socio-ecological systems and go on to discuss measures to treat these limitations and design inclusive, reliable and responsible computer simulations for achieving sustainable development.

Seán Ó hÉigeartaigh, John Burden and Jose Hernandez-Orallo were co-authors of a paper examining capabilities and safety challenges associated with language models, published in the Proceedings of the AAAI 2022.

How General-Purpose Is a Language Model? Usefulness and Safety with Human Prompters in the Wild in *Proceedings of the AAAI Conference on Artificial Intelligence* by Pablo Antonio Moreno Casares, Bao Sheng Loe, John Burden, Seán Ó hÉigeartaigh, and Jose Hernandez-Orallo, 22 June 2022.

> The new generation of language models is reported to solve some extraordinary tasks the models were never trained for specifically, in few-shot or zero-shot settings. However, these reports usually cherry-pick the tasks, use the best prompts, and unwrap or extract the solutions leniently even if they are followed by nonsensical text. In sum, they are specialised results for one domain, a particular way of using the models and interpreting the results. In this paper, we present a novel theoretical evaluation framework and a distinctive experimental study assessing language models as general-purpose systems when used directly by human prompters in the wild. For a useful and safe interaction in these increasingly more common conditions, we need to understand when the model fails because of a lack of capability or a misunderstanding of the user's intents. Our results indicate that language models such as GPT-3 have limited understanding of the human command; far from becoming general-purpose systems in the wild.

4.2 Preprints

Matthijs Maas and Kayla Lucero-Matteucci published a preprint on military applications of AI, arguing that GCR researchers may be focusing too much on the threat from Lethal Autonomous Weapons and not enough on the applications of AI to nuclear weapons systems.

Military Artificial Intelligence as Contributor to Global Catastrophic Risk SSRN pre-print 27 May 2022 by Matthijs Maas, Kayla Lucero-Matteucci, Di Cooke.

> Recent years have seen growing attention for the use of AI technologies in warfare, which has been rapidly advancing. This chapter explores in what ways such military AI technologies might contribute to Global Catastrophic Risks (GCR). After reviewing the GCR field's limited previous engagement with military AI, and giving an overview of recent advances in military AI, this chapter focuses on two risk scenarios that have been proposed. First, we discuss arguments around the use of swarms of Lethal Autonomous Weapons Systems, and suggest that while these systems are concerning, they appear not yet likely to be a GCR in the near-term, on the basis of current and anticipated production limits and costs which make these systems still uncompetitive with extant systems for mass destruction. Second, we delve into the intersection of military AI and nuclear weapons, which we argue has a significantly higher GCR potential. We review historical debates over when, where, and why nuclear weapons could lead to GCR, along with recent geopolitical developments that could raise these risks further. We then outline six ways in which the use of AI systems in-, around-, or against- nuclear weapons and their command infrastructures could increase the likelihood of nuclear escalation and global catastrophe. The chapter concludes with suggestions for a research agenda that can gain a more comprehensive and multidisciplinary understanding of the potential risks from military AI, both today and in the future.

Lalitha Sundaram, Matthijs Maas, and SJ Beard published a preprint exploring current issues in the field of Existential Risk Studies.

Seven Questions for Existential Risk Studies SSRN preprint June 7 2022 by Lalitha Sundaram, Matthijs Maas, SJ Beard.

Recent years have seen the emergence of Existential Risk Studies (ERS), a rich field focused on understanding and mitigating a range of Extreme Technological Risks. This interdisciplinary and idiosyncratic field today finds itself at a crossroads: at the same time as many risks are growing increasingly urgent, there is increasing attention and potential to shape global action in response. The magnitude of both risks and opportunities create an urgent need for this field to reflect on how it defines itself going forward–as a community, as a science, and as a project pursuing change. Drawing on work from across the ERS field and beyond, as well as the authors' personal experiences, this article seeks to spur and aid this process of reflection. To that end, we pose and discuss seven key questions for the ERS field. These are: (1) how can scholars of ERS choose which risks to prioritize? (2) How or why can we prioritize extreme technological risk over other global problems, and what are the possible downsides of ERS? (3) What are the different possible approaches to studying and managing ETRs? (4) In what ways can the field of ERS pursue a coherent scientific approach? (5) How can or should ERS pursue impact to mitigate ETRs? (6) How diverse is the field of ERS, and how can greater diversity aid the ERS field? (7) How can ERS best reflect on -and communicate about these questions, as it continues to grow? We do not provide definitive answers, but attempt to explore how and why different people might respond in different ways. We point to open debates around: disciplinary direction, pluralism, epistemic modesty, diversity, inclusion, representation and accountability. We take these challenges to be not just important but increasingly urgent, as pivotal considerations in charting a path forward for our field–and, perhaps, our world.

4. Appendix Summary of the 4th Cambridge Conference on Catastrophic Risk

Day 1. Future Risks and how to study them

The day opened with addresses by CSER's co-founder, Martin Rees, providing personal reflections on existential risk and the work of the centre, and CSER's Deputy Director, Jess Bland, who talked about 'Strange Aids to Thought: Why Do Imagined Worlds Help Build Resilience to Very Real Catastrophic Risks?'

After a short ice breaker exercise from Pablo Suarez of the Red Cross to introduce participants both in the room and on-line, the day continued with a presentation on CSER's approach to foresight and horizon scanning by Luke Kemp. Luke presented different foresight and forecasting methods for global catastrophe: their strengths, limitations, and how they can be combined to create a more accurate and robust foresight system.

The next session focused on different perspectives on how to study future risks from Artificial Intelligence, with a particular focus on techniques that fully account for the diversity of contexts surrounding its development. This was opened with a keynote address by Tina Park from the Partnership for AI on Addressing the Challenges of Inclusive Practises in AI Development.

The discussion was then continued by a panel of Jess Whittlestone, from the Centre for Long Term Resilience, who talked about the analytical exploration of future risks, Shahar Avin, from CSER, who talked about exploring extreme risks through role-play, and David Krueger, from the Department for Engineering, who talked about 'Deep Learning, Scaling, Alignment and Existential Risk'.

The afternoon opened with a split programme with some in-person participants joining a workshop on foresight and horizon scanning lead by Luke Kemp while online and in-person attendees could also enjoy a series of lightning talks on a wider variety of topics around existential and global catastrophic risk:

- Ross Tieman, 'Digital Fragility – Digitization of Critical Infrastructure and Increased Risk of Catastrophic Failures'
- Caroline Baylon, 'Tackling Interconnected Global Risks: The Need for Long-term Thinking and a Multilateral Approach'
- Nora Ammann, 'Learning from Existing Complex System about Existential Risks and Alignment'
- Markus Reichstein, 'Existential risk – Emerging from systemic and compound risk?'
- Eamon Aloyo, 'The Catastrophic Risk Reduction Case for Funding Research on Stratospheric Aerosol Removal'
- Michael Cassidy, 'Large magnitude volcanic eruptions as global catastrophic risks and existential risk factors'
- Anders Sandberg, 'Volcano engineering ethics'
- Felix Riede, 'Apocalypse then, Apocalypse now? Building realistic disaster scenarios for low-frequency/high-magnitude volcanism in Europe using the Laacher'

- Matthew Rendall, ‘Nuclear war as a predictable surprise’
- David Denkenberger, ‘Integrated assessment of food production in response to global catastrophic food risks’
- Alix Pham, ‘Balanced diets on resilient foods in abrupt sunlight reduction scenarios’
- Dennis M Bushnell, ‘Halophytes For Land, Water, Food, Energy, Climate’

Finally, all participants came back together for a closing address by Joachim Isacsson titled “Tomorrow Never Dies: Bolts from the Blue and Creeping Crises, Disruptions in a Changing World”, which drew on his experiences working for the Development, Concepts, and Doctrine Centre at the UK Ministry of Defence.

Day 2. Real Catastrophes and what we can learn from them

The day opened with a talk by one of CSER’s academic programme managers, SJ Beard, on the importance of thinking about real catastrophes in relation to existential risk and how we can explore these through maps and stories. Lara Mani then gave a presentation on CSER’s approach to outreach and communications. Her presentation explored what we can learn from real-time disasters for communicating risk and how we can engage stakeholders and publics with information about GCRs. It also highlighted the importance of evaluation for successful communication.

The next session reflected on a variety of historical catastrophes with relevance to the study and mitigation of GCR, with a particular focus on how public discourse about these disasters often takes a narrow focus and learns the wrong lessons. It was opened by Robin Gorna, an independent writer and activist working on public health and social justice, who talked about lessons across pandemics, from AIDS to COVID.

The discussion was then continued by a panel who focused on a variety of catastrophes, including Julius Weitzdörfer, from Hagen University, talking about lessons from Fukushima, Lalitha Sundaram, from CSER, talking about AIDS and other chronic diseases as GCRs, and Jochem Rietveld, also from CSER, talking about our emerging Lessons from COVID-19 project.

Before breaking for lunch, Nandini Shiralkar gave a talk about her work to establish the Cambridge Existential Risk Initiative and how researchers could engage with its projects.

Once again, the afternoon opened with a split session featuring an in-person workshop on scenarios run by Lara Mani and a series of on-line lightning talks including:

- Kayla Lucero-Matteucci, ‘The Command and Control of Nuclear Weapons Under Increased Pressure’
- Anders Sandberg, ‘A Safe Governance Space for Humanity: Necessary Conditions for the Governance of Global Catastrophic Risks’
- Simeon Campos, ‘What Travel Networks Can Teach Us about Future Pandemics and GCBRs’
- Rumtin Sepasspour, ‘The policy relevance of the existential risk studies field’
- Aaron Tang, ‘A Fate Worse Than Warming? Stratospheric Aerosol Injection and Global Catastrophic Risk’
- Daniel Bertram, ‘Ecocide: Can International Criminal Law Prevent Ecological Collapse?’
- Nathaniel Cooke, ‘Weathering the Storm: Societal Resilience to Existential Catastrophes’
- Murilo Karasinski, ‘To make it more complex: axiological futurism in the reflection on existential risks’
- Shira Ahissar, ‘The risk of a superintelligence and the Precautionary Principle’

- Bear Häon, 'Deceptive AI: A Blueprint for Legal and Technical Synergy'
- Nick Wilson, 'Catastrophe, X-risk and preserving island nodes of complexity'
- Matt Boyd, 'Step 1 in solving existential risks: include them in national risk assessments'

Ariel Conn then gave a short talk about her work to establish Technology, Arts and a New Global Objective for the Future (TANGO Future) a project aimed at bringing more diverse conversations into discussions about future technologies and global challenges.

Finally, the day closed with an address by Bryan Walsh, editor of *Vox*'s Future Perfect vertical on 'Reporting On the End of the World: The Challenge of Covering Long-Term Risks in a Short-Term Media World', in which he talked about his future plans and how to make good editorial decisions about the issues that matter most.

Day 3. Global Solutions and how we can implement them

The day opened with a talk by one of CSER's academic programme managers, Paul Ingram, drawing on 30 years working on nuclear disarmament and asking how can we best respond to existential risk? This was then followed by a talk from Clarissa Rios Rojas about CSER's Approach to Policy. This outlined the co-creation of policy and research with academics and policy brokers through a GCR Science-Policy Interface group and the use of science diplomacy.

The next session reflected on the range of institutions that have a role to play in preventing global catastrophes and the challenges they face in doing this, with a particular focus on overcoming barriers to global cooperation and cross-cultural understanding. It opened with a keynote address by Jenty Kirsch-Wood, of the United Nations Office for Disaster Risk Reduction, talking about UNDRR's Preparation for Systemic and Cascading GCRs.

The discussion was then continued by a panel who focused on bringing concerns about global catastrophic and existential risk into a range of policy forums including, Shin-Shin Hua, a competition and tech lawyer, who talked about the regulation of AI: governing the ungovernable?, James Ginns, of the Centre for Long Term Resilience, who talked about private sector perspectives on risk management in government and its global application, and Max Stauffer, of the Simon Institute, who talked about engaging with global institutions.

The conference formally closed with a talk by Oliver Letwin, former UK cabinet minister, on 'Planning for Catastrophe: Why Resilience Equals Fallback'. This drew upon his experience of being responsibility for national resilience and civil contingencies for the UK government and his subsequent work on risk planning and public policy. Following this, in person attendees were able to attend a third workshop run by Clarissa Rios Rojas on her developing Science Policy Interface for Global Catastrophic Risk.

Contact

Elizabeth Brent
Associate Director — Arts and Humanities
University of Cambridge Development and
Alumni Relations
1 Quayside, Bridge Street
Cambridge
CB5 8AB
Elizabeth.Brent@admin.cam.ac.uk
+44 (0)1223 762891
www.cam.ac.uk/yourscambridge

Dr SJ Beard
Academic Programme Manager and
Senior Research Associate
Centre for the Study of Existential Risk
16 Mill Lane
Cambridge
CB2 1SB
sjb316@cam.ac.uk
(+44) 01223 766838
www.cser.ac.uk

UNIVERSITY OF CAMBRIDGE

Dear World... Yours, Cambridge

The campaign for the University and Colleges of Cambridge