

Centre for the Study of Existential Risk

MARCH 2024



Contents

An introduction from Matt Connelly

Overview

People

1.1 New Members of Staff

1.2 Visitors

1.3 Research Affiliates

1.4 Leavers

Events, Engagement and Outreach

2.1 Academic engagement

2.2 Policy and Stakeholder Engagement

2.3 Public and Media Engagement

Publications

3.1 Papers

3.2 Books

3.3 Reports


Contact

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.

Return to contents

University of Cambridge



3
4
6
6
6
6
7
8
8
9
10
12
12
17
17
20

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.

An introduction from Matt Connelly CSER director

In this second termly report since I became CSER director, I thought it would be useful to summarise the strategy I have been working to develop.

This strategy starts with the premise that catastrophic and existential risks are serious and persistent, and mitigation requires basic and applied multidisciplinary research together with ongoing policy engagement, both of which must be sustained over the long run.

Recent events show the importance of embedding this effort in institutions and institutional arrangements that can survive fluctuations in funding and popular interest. Social movements come and go, but universities endure.

I have pursued a multi-pronged strategy for the long-term growth to make CSER an integral and visible part of Cambridge, and earn recognition of catastrophic and existential risk studies as a thriving field of research in the academy. This includes:

1) Forming a broad coalition behind basic principles for why we do this work, and why it matters. CSER will soon make available a public statement of 'Essential Principles'.

2) Institutionalising CSER at the centre of this field. The statement will be released at the same time we announce our fifth bi-annual conference.

3) Preparations are also well underway for an MPhil programme, the first in the field. The programme will ensure the positions of at least two senior teaching roles and administrative support. At the same time, we should begin to see the fruits of a massive expansion in applications for external funding: autumn 2023 sees a 10-fold increase over the corresponding 6-month period in 2022.

4) Continuing efforts to secure large gifts from individuals and family foundations, pursued in collaboration with other Cambridge centres (for example, CRASSH, Cambridge Zero, the Centre for Geopolitics).

5) Increasing the visibility of CSER, now supported by a new full-time position in communications. In addition to the conference, we continue to organise public talks (like Yuval Harari's, in February 2024) and I will be launching an eight-part podcast ('The History of the End of the World') that will feature Yuval, Martin Rees, and Niall Ferguson, among others. We also sent out a mass mailing of letters and brochures to



hundreds of people who signed last May's statement on AI-risk.

6) Policy engagement, including four different secondments, which has allowed engagement with multiple UK ministries. I have also continued my own efforts to engage the U.S. House and Senate on strengthening legal requirements for transparency and accountability. This contributed to the first-ever legal requirement – as part of the 2024 National Defense Authorization Act – that the executive branch develop a plan to use AI to better manage national security information.

I look forward to discussing and developing these ideas with many of you during 2024.

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 civilisational 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 outlines our activities from September to December 2023. It is produced initially as a paper to support our Strategy Group in reviewing and planning CSER's research and impact, and shared with other governance bodies within the University of Cambridge. It is then designed into a report, shared with stakeholders and funders, as well as on the CSER website to maximise our transparency about the team's activities and outputs.

Highlights of the last three months include:

- CSER researchers developed closer relationships with disciplines and academic communities that we had been less engaged with previously, including complex system modellers gathered at the Max Planck Institute for the Physics of Complex Systems in Dresden, ancient historians and environmental scientists from Malaysia, Nigeria and India, and with colleagues in a range of North American institutions from John Hopkins University and the Roosevelt Institute for American Studies to Balsillie School of International Affairs in Canada.
- Several CSER and CFI researchers produced work and took up government secondments that were influential in framing the agenda for the UK AI Summit in late October. Key resources related to this include Shahar Avin's paper and Seán Ó hÉigeartaigh's commentary afterwards. In late November and early December researchers travelled to international governance meetings for nuclear and biological risks, including Alex Klein winning an NTI competition and presenting a paper on AI and Biological risk at the Biological Weapons Convention Meeting of States Parties.

Martin Rees at the Alumni Festival panel discussion on 23 September.



- This period saw significant media interest in Lara Mani's paper on the ethics of volcanic engineering, as well as a sold-out session at the Cambridge Alumni Festival that explored the themes of The Era of Global Risk through three researchers' projects.
- CSER and affiliates produced 10 reviewed papers and 4 pre-prints, covering a wide range of topics from qualitative and quantitative research methods to conceptual concerns about policy, ethics and an analysis of how socio-ecological systems collapse, as well as a number of papers on environmental risks including oceans, food and volcanic interventions.
- Recent CSER alumnus Lauren Holt produced a rich publication, blending commentary and creative writing, along with an audio artwork and illustrations that were produced as part of the project. The works are available on the CSER website.
- Lara Mani and Lalitha Sundaram co-produced an influential report with the United Nations Office for Disaster Risk Reduction (UNDRR) and other collaborators. Gideon Futerman and SJ Beard produced a comprehensive summary of a workshop on the relationships between solar radiation management technologies and catastrophic risk. CSER Affiliates were involved in two policy reports.



Alex Klein and colleagues won NTI competition and presented their paper at BWC.

People

1.1 New Members of Staff



[Ken Mbeva](#) previously served as a Postdoctoral Research Associate at the Blavatnik School of Government, University of Oxford. Dr Mbeva's research examines how risk and uncertainty shape international

cooperation and global governance, focusing on economic, sustainability, and demographic issues. In his current role, he studies how to manage the global risks generated by rapid demographic changes, thus averting societal collapse and existential risks to humanity.

Dr Mbeva has contributed to high-profile UN scientific reports such as the IPCC AR6 and the Adaptation Gap Report and served on Kenya's official delegation to the UN Climate Change negotiations. A Kenyan national, he holds a PhD in International Relations (distinction) from the University of Melbourne, and has studied and worked in several continents, including Africa, Europe, Asia, and Australia. His latest book is *Africa's Right to Development in a Climate-Constrained World* (Palgrave Macmillan, 2023, co-authored).

1.2 Visitors

We have welcomed one new visitor during this period:



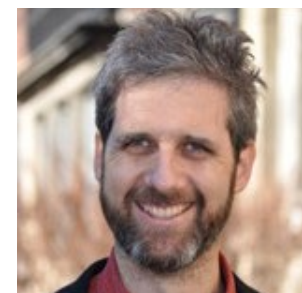
[Francesco Bertolotti](#) is a postdoctoral researcher who specialised in designing dynamical models of complex systems for problem-solving and prediction activities, with a particular emphasis on agent-

based models. His primary research focuses on utilising simulations and agent-based modelling to uncover the underlying mechanisms driving changes in risk preferences. This work aims at contributing to a better understanding of how different entities make decisions in non-trivial environments.

At CSER, he is working on a sustainability game to better understand how short-term decisions can affect the collapse of societal systems when resources are limited and have a fixed renewal time. Moreover, they plan to use the time at CSER to write a commentary on the relationship between sustainability and system science.

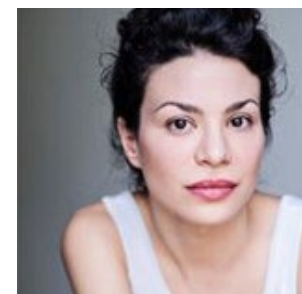
1.3 Research Affiliates

We have welcomed four new research affiliates.



[Pablo Suarez](#) is a system dynamics modeller 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.



[Taniel Yusef](#) is a researcher and advocate working on various portfolios at the UN, European and UK parliaments, with an emphasis on disarmament. She specifically focuses

on weapons technology regulation, (AI, nuclear weapons, outer space threats and cyber-security), as well as supply chain, global trade, development, emerging economies and gender. Taniel has a number of roles including Technology Developers Coordinator of the UK Campaign to Stop Killer Robots, visiting lecturer at the University of East London and WILPF Advisory Board, recently stepping down as WILPF International Representative (UK). She researches conflict and resilience on the ground, particularly with a gendered lens.



Cecil Abungu is a PhD student at the University of Cambridge, coordinator of the ILINA Program and Research Affiliate with the Legal Priorities Project. Cecil also serves as an advisor at the AI Futures

Fellowship and lecturer at Strathmore Law School. Cecil's recent research has focused on the role of developing countries in global catastrophic risk-related governance of AI, traditional African thought on the long-term future and algorithmic discrimination. He holds an undergraduate law degree from Strathmore Law School in Nairobi and Master's in law degree from Harvard Law School.



Internet Hero of the Year' by the Internet Service Providers' Association, and was Chair of the Panel of Independent Reviewers for DeepMind Health. He is a Director of the Joseph Rowntree Reform Trust Ltd and has advised many organisations in the public and private sectors.

Dr Julian Huppert is Founding Director of the Intellectual Forum in Jesus College, Cambridge, and was the Member of Parliament for Cambridge. He is noted for his work in technology policy, being named '2013

1.4 Leavers

We have sadly said goodbye to two of our researchers, who will remain CSER Research Affiliates.

- Alex McLaughlin took up a permanent position as lecturer in Global Political Theory at the University of Exeter. He is joining the Centre for Political Thought, one of the largest clusters of theorists and historians of political thought in the UK. Alex's skills and interest in fostering open, respectful debate among diverse views will be missed, but this is a fantastic opportunity to join an institution that was founded on that very idea.
- Sabin Roman continues to publish as a CSER Affiliate (with two excellent papers this term) while he sets up a new organisation called the Odyssean Institute, founded to incorporate robust decision making and deliberative mechanisms, both expert and public, with existential risk mitigation.

Events, Engagement and Outreach

2.1 Academic engagement

CSER researchers developed closer relationships with disciplines and academic communities that we had been less engaged with previously, including complex system modellers gathered at the Max Planck Institute for the Physics of Complex Systems in Dresden, ancient historians and environmental scientists from Malaysia, Nigeria and India, and with colleagues in a range of North American institutions from John Hopkins University and the Roosevelt Institute for American Studies to Balsillie School of International Affairs in Canada.

- 9 October: the [RISK-KAN](#) working group 'Learning from the past for the future', of which Lara Mani is a member, hosted a webinar on 'How monsoon and megadrought shaped the rise and fall of a Harappan city Dholavira' from Prof Anindya Sarkar (Indian Institute of Technology Kharagpur)
- 16 October: Constantin Arnscheidt and Sabin Roman attended a workshop on 'Non-autonomous Dynamics in Complex Systems: Theory and Applications to Critical Transitions' at the Max Planck Institute for the Physics of Complex Systems in Dresden. Constantin gave a talk on 'Rate-induced tipping and global catastrophic risk'. Sabin gave a talk on 'Global history, the emergence of chaos and inducing sustainability in networks of socio-ecological systems'
- 23 October: Matthew Connelly held a lecture for the Roosevelt Institute for American Studies Lecture titled 'America's Secrecy Industrial Complex: History and the Future'
- 24 October: Lalitha Sundaram spoke to the Biosecurity Salon, organised by Johns Hopkins University and MIT
- 26 October: The first session of the [CRASSH Healthcare in Conflict Network](#) that is co-convened by CSER Affiliate Charlotte Hammer
- 12 November: Matthew Connelly presented [Data for Good](#) for the Columbia Data Science Institute
- 22 November: Constantin Arnscheidt gave an invited lecture at the School of Computing and Mathematical Sciences, University of Leicester. Talk entitled 'Nonlinear Earth system dynamics: stability and catastrophe'
- 11 December: Constantin Arnscheidt and Lara Mani attended the Mimir Center launch and Lara gave a talk on X-risk Communication



Lara Mani at the Mimir Center launch.

- 17 December: Constantin Arnscheidt presented at the Critical Transitions Workshop IV at the Earth Resilience and Sustainability Initiative on 'Critical transitions and global catastrophic risk'
- 18 December: Lara Mani gave a talk at the UK Alliance for Disaster Research (UKADR) conference on 'A transdisciplinary approach for improving global preparedness for large magnitude volcanic eruptions'

2.2 Policy and Stakeholder Engagement

Several CSER and CFI researchers produced work and took up government secondments that were influential in framing the agenda for the UK AI Summit in late October. Key resources related to this include Shahar Avin's paper and Seán Ó hÉigeartaigh's commentary afterwards. In late November and early December researchers travelled to international governance meetings for nuclear and biological risks, including Alex Klein winning an NTI competition and presenting a paper on AI and Biological risk at the Biological Weapons Convention Meeting of States Parties.

- 14 September: Lalitha Sundaram attended her first meeting as a member of the UK's Biosecurity Leadership Council
- 19 September: Lara Mani and Lalitha Sundaram joined a meeting of the ASRA (Accelerating Systemic Risk Assessment, supported by the UN Foundation) Working Group on 'State of the Practice'
- 3 October: Lalitha Sundaram met Scott Janzwood, Michael Lawrence and Megan Shipman from the Cascade Institute
- 5 October: Jochem Rietveld met Jasmin Kaur from 1 Day Sooner to explore potential shared outreach activity
- 31 October: The framework set out in Shahar Avin's earlier paper with colleagues at OpenAI, Deepmind, Anthropic and others was influential on the agenda for the UK AI Safety summit, as was the secondment from earlier in the year
- 26 October: CSER co-founder Jaan Tallinn and CFI Associate Fellow Yi Zeng were appointed to the [UN High-level Advisory Body on Artificial Intelligence](#)
- 1 November: CSER provided written input to the UK Government on the importance of research into high-impact risks, an area subsequently funded with £10m in the Autumn Statement
- 17 November: Lara Mani met with commissioners from the National Infrastructure Commission for Wales to discuss existential risk communication
- 27 November – 1 December: Paul Ingram attended the meeting of the Treaty on Prohibition of Nuclear Weapons at the UN, New York, consulting with several states parties to discuss contemporary approaches to nuclear disarmament diplomacy in the context of conflict. Paul also chaired a side event on a Middle East Treaty for WMD on 1 December
- 30 November: A team including CSER's Alex Klein, Gurpreet Dhaliwal and Askar Kleefeldt won the [2023 Next Generation for Biosecurity Competition](#). The team received travel support to attend the Biological Weapons Convention Meeting of States Parties in Geneva, Switzerland to present their paper 'The Convergence of AI and the Life Sciences: Safeguarding Biotechnology, Bolstering Biosecurity, and Supporting Bioeconomies' during an NTI side event
- 5 December: Maurice Chiodo gave a talk at a Sarasin AI Investor meeting
- 10 - 14 December: Paul Ingram attended a [Geneva Centre for Security Policy meeting](#) with participants from the nuclear five, alongside consultations at missions on approaches to nuclear disarmament diplomacy
- 12 December: Matt Connelly mailed more than 450 of the signatories of the May 2023 statement that mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war, outlining CSER's work and asking if they would be interested in working together
- 17 December: Tom Hobson, Haydn Belfield, Alex Klein and Lalitha Sundaram attended a Harvard Sussex Programme, BASIC and UK Foreign, Commonwealth & Development Office workshop on AI and Biosecurity

2.3 Public and Media Engagement

This period saw significant media interest in Lara Mani's paper on the ethics of volcanic engineering, as well as a sold-out session at the Cambridge Alumni Festival that explored the themes of The Era of Global Risk through three researchers' projects.

- 7 September: CSER and CFI alumni Jess Whittlestone featured in [Time Magazine's 100 most influential people in AI](#)
- 14 September: Lara Mani, John Burden and Henry Shevlin spoke to the [Future of Life Institute podcast](#) about their FHI Worldbuilding Contest 'Core Central'
- 17 September: BBC Radio 4 broadcast CSER visitor Sarah Woods' drama-documentary inspired by Tocqueville's influential survey of democracy, *Tocqueville's Democracy in America*
- 20 September: Matt Connelly chaired a Cambridge Zero panel as part of the UN Climate Week in New York on the theme of creating a better planetary future
- 23 September: CSER [hosted a panel discussion](#) with Martin Rees, Jessica Bland, Lalitha



September 23 panel discussion with, from left, Jessica Bland, Constantin Arnscheidt, Lalitha Sundaram and Haydn Belfield.

Sundaram, Haydn Belfield and Constantin Arnscheidt as part of the Cambridge Alumni Festival about CSER's new book, *The Era of Global Risk*

- 2 October: Matt Connelly published an extract from his recent book as an article in Literary Hub called [How US Intelligence Agencies hid their most shameful experiments](#)
- 3 October: Lara Mani was [interviewed by the](#)

[Daily Mail](#) about the impact of volcanic events

- 24 October: Matt Connelly gave a [talk on Declassification](#) to the Cambridge Existential Risk Initiative (a Cambridge University student society)
- 24 October: [HowTheLightGetsIn festival published a video](#) of Lara Mani's co-author Anders Sandberg talking about their paper on the ethics of volcano geoengineering

- 7 November: Lara Mani gave a talk on [building societal preparedness for globally disruptive volcanic eruptions](#) organised by the Cambridge University Science and Policy Exchange
- 8 November: Seán Ó hÉigeartaigh wrote an [article for the Cambridge University website](#) about the UK AI Summit
- 20 November: Haydn Belfield [spoke on a panel](#) organised by the Edinburgh Futures Institute about the geopolitics of AI
- 13 December: Affiliate Mike Cassidy [joined the Reviewer 2 podcast](#) to discuss the paper he co-authored with Lara Mani and Anders Sandberg
- 21 December: CSER affiliate Mike Cassidy spoke to [BBC Futures](#) about [volcanic geoengineering](#).



Top right, Lara Mani gives a talk at CUSPE on 7 November.

Bottom right, Haydn Belfield, right, at the Edinburgh Futures Institute panel on November 20, with, from left, John Zerilli, Kate Kaye and Kerry McNerney.



Publications

3.1 Papers

CSER and Affiliates produced 10 reviewed papers and 4 pre-prints, covering a wide range of topics from qualitative and quantitative research methods to conceptual concerns about policy, ethics and an analysis of how socio-ecological systems collapse, as well as a number of papers on environmental risks including oceans, food and volcanic interventions.

[ParEvo: A methodology for the exploration and evaluation of alternative futures](#) in *Evaluation*
Rick Davies, Tom Hobson, Lara Mani, SJ Beard
10 September 2023

Evaluators' main encounter with views of the future is in the form of theories of change, about how a programme will work to achieve a desired end, in a given context. These are typically focussed on specific relatively short-term futures, which are both desired and expected. But even in the short term, reality often involves unpredictable events which must be responded to. Other ways of thinking about the future may be helpful and complementary, notably those developed by foresight practitioners working in the field of futures studies. These pay more attention to a range of possible futures, rather than a single perspective. One way of exploring such

futures is by using ParEvo.org, an online process that enables the participatory exploration of alternative futures. This article explains how the ParEvo process works, the theory informing its design, and its usage to date. Attention is given to three evaluation challenges, and methods to address them: (a) optimising exercise design, (b) analysis of immediate results and (c) identifying longer-term impacts. Two exercises undertaken by the Cambridge-based Centre for the Study of Existential Risk (CSER) in 2021–2022 are used as illustrative examples.

[Climate Resistance and the Far Future](#) in *Social Theory and Practice*
Alex McLaughlin
23 September 2023

This paper argues that climate injustice will be compounded in the future as a result of the deferred nature of many climate impacts. My claim is that the temporal disconnect between emissions and climate harm threatens future people's ability to access what I call "resistance goods," which rely on forms of address, often realised in oppositional political action. I identify three resistance goods—self-assertion, solidarity and testimony—and show that each is threatened by the temporality of climate change. A compound of climate injustice is that it will be experienced as demeaning, isolating and silencing by many future people.

[Nathan Sears: "... in the midst of catastrophe"](#) in *Global Policy*
Haydn Belfield
25 September 2023

Nathan Sears had begun a strikingly ambitious intellectual project of reimagining international relations and world order, almost from the ground up.

Nathan saw that we were in a new era of human history, an era of existential risk. We face threats of global civilisation collapse, possibly even human extinction, of our own making. These anthropogenic risks include nuclear weapons, biological weapons—especially new engineered pandemics—environmental risks such as climate change and risks from emerging technologies, such as artificial general intelligence (AGI).

As Nathan saw it, this new era should pose profound challenges to traditional ways of thinking about and doing international relations (IR; Sears, 2020a). Instead of 'human security' of individual citizens or 'national security' of individual states, we must now also be concerned with 'existential security' of the whole of humanity. Instead of thinking of army divisions or status games when ascertaining Great Powers, we must now ask, 'who can destroy the world?'. The old Westphalian order

struggles to make sense of this new order. The old units of states and alliances, the old methods of war and balance of power and the old ideologies of self-help and power politics are inadequate for comprehending this new era. Nathan's work took seriously the challenge existential risks pose to existing theories of international relations. He gazed directly into the abyss of civilisational collapse and human extinction and sought to reimagine rationalism, neorealism and securitisation theory within this frame.

[Existential security: Safeguarding humanity or globalising power?](#) in *Global Policy*

Tom Hobson and CSER Affiliate Olaf Corry
25 September 2023

Nathan Sears' (2020) exploration of how a policy of 'existential security' might be fostered represents one of the first efforts to systematically think through security and how it might relate to thinking about existential risks.

The concept of existential risk emerged in the early twenty-first century (see, e.g. Bostrom, 2002). It refers to the idea that there are a class of hazards which may 'threaten the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development' (Bostrom, 2013, p. 15). As a new field of study, Existential Risk Studies (ERS) is small but quickly

expanding. A number of research centres have attracted significant attention – from both the media and policymakers – and large amounts of funding from high-profile private individuals and philanthropic foundations. The concept is also entering mainstream political discourse. In recent months, UK Prime Minister Rishi Sunak has met with leaders of industry in AI research to discuss existential risks¹ and references to their existence have proliferated in both the vernacular of political elites and in policy reports and white papers.

But how all this relates to the concept and practices of security is by no means straightforward.

[Long-term Feedback Mechanisms Underlying Societal Growth and Collapse](#) an SSRN pre-print

Sabin Roman
2 October 2023

In this work we address some common methodological pitfalls in understanding societal collapse and propose a framework to remediate them. With this goal we reformulate and extend Tainter's theory of societal collapse to a dynamical setting by identifying the feedback mechanisms that underlie it and how these can be applied to specific historical cases. We find a archetypal pattern of feedback loops that relate different measures of complexity, of resource exploitation and of generated returns that

characterize a society. The general pattern, called the theta process, is instantiated by proposing specific feedback relationships that operated in different societies over timescales of centuries, accounting for both long-term growth and eventual decline or collapse. The framework allows us to classify societies depending on the key activities and institutions they rely on: agriculture for Easter Island and the Lowland Classic Maya, the military and state bureaucracy for the Western Roman Empire and Imperial China, religious beliefs for the Egyptian Old Kingdom and the Greenland Norse and trade activity for the East Mediterranean basin during the Bronze Age. Overall, the framework provides a unifying perspective on a wide diversity of historical cases of collapse and proposes feedback mechanism analysis to be at the forefront of conceptualizing societal dynamics.

[Predictable Artificial Intelligence](#) on *arXiv*

John Burden, Alex Marcoci and Seán Ó hÉigeartaigh from CSER with José Hernández-Orallo, CSER Affiliate and Lexin Zhou, Pablo A. Moreno-Casares, Fernando Martínez-Plumed, Ryan Burnell, Lucy Cheke, Cèsar Ferri, Behzad Mehrbakhsh, Yael Moros-Daval, Danaja Rutar, Wout Schellaert and Konstantinos Voudouris
9 October 2023

We introduce the fundamental ideas and challenges of Predictable AI, a nascent research area that explores the ways in which we

can anticipate key indicators of present and future AI ecosystems. We argue that achieving predictability is crucial for fostering trust, liability, control, alignment and safety of AI ecosystems, and thus should be prioritised over performance. While distinctive from other areas of technical and non-technical AI research, the questions, hypotheses and challenges relevant to Predictable AI were yet to be clearly described. This paper aims to elucidate them, calls for identifying paths towards AI predictability and outlines the potential impact of this emergent field.

[Digital twins: a stepping stone to achieve ocean sustainability?](#) in *NPJ Ocean Sustainability*
Asaf Tzachor, Catherine Richards, CSER Affiliates with Ofir Hendel
9 October 2023

Digital twins, a nascent yet potent computer technology, can substantially advance sustainable ocean management by mitigating overfishing and habitat degradation, modeling, and preventing marine pollution and supporting climate adaptation by safely assessing marine geoengineering alternatives. Concomitantly, digital twins may facilitate multi-party marine spatial planning. However, the potential of this emerging technology for such purposes is underexplored and yet to be realized, with just one notable project entitled European Digital Twins of the Ocean. Here, we consider the

promise of digital twins for ocean sustainability across four thematic areas. We further emphasize implementation barriers, namely, data availability and quality, compatibility, and cost. Regarding oceanic data availability, we note the issues of spatial coverage, depth coverage, temporal resolution, and limited data sharing, underpinned, among other factors, by insufficient knowledge of marine processes. Inspired by the prospects of digital twins, and informed by impending difficulties, we propose to improve the availability and quality of data about the oceans, to take measures to ensure data standardization, and to prioritize implementation in areas of high conservation value by following the ‘nested enterprise’ approach.

[Scoping Potential Routes to UK Civil Unrest via the Food System: Results of a Structured Expert Elicitation](#) in *Sustainability*
SJ Beard from CSER and Asaf Tzachor, CSER Affiliate with many others
12 October 2023

We report the results of a structured expert elicitation to identify the most likely types of potential food system disruption scenarios for the UK, focusing on routes to civil unrest. We take a backcasting approach by defining as an end-point a societal event in which 1 in 2000 people have been injured in the UK, which 40% of experts rated as “Possible (20–50%)”, “More

likely than not (50–80%)” or “Very likely (>80%)” over the coming decade. Over a timeframe of 50 years, this increased to 80% of experts. The experts considered two food system scenarios and ranked their plausibility of contributing to the given societal scenario. For a timescale of 10 years, the majority identified a food distribution problem as the most likely. Over a timescale of 50 years, the experts were more evenly split between the two scenarios, but over half thought the most likely route to civil unrest would be a lack of total food in the UK. However, the experts stressed that the various causes of food system disruption are interconnected and can create cascading risks, highlighting the importance of a systems approach. We encourage food system stakeholders to use these results in their risk planning and recommend future work to support prevention, preparedness, response and recovery planning.

[Teaching Resources for Embedding Ethics in Mathematics: Exercises, Projects, and Handouts](#) on *arXiv*
Maurice Chiodo from CSER with Dennis Müller
12 October 2023

The resources compiled in this document provide an approach to embed and teach Ethics in Mathematics at the undergraduate level. We provide mathematical exercises and homework problems that teach students

ethical awareness and transferable skills, for many of the standard courses in the first and second years of a university degree in mathematics or related courses with significant mathematical content (e.g., physics, engineering, computer science, economics, etc). In addition to the exercises, this document also contains a list of projects, essay topics, and handouts for use as final projects and in seminars. This is a living document, and additional contributions are welcome.

The Ethics of Volcano Geoengineering in Earth's Future

Lara Mani from CSER and Mike Cassidy, CSER Affiliate with Anders Sandberg
20 October 2023

Volcano geoengineering is the practice of altering the state of volcanic systems and/or volcanic eruptions to exploit them or mitigate their risk. Although many in the field insist there is little that can be done to mitigate the hazard, past examples of both intentional and inadvertent volcano interventions demonstrate that it is technically feasible to reach volcano plumbing systems or alter atmospheric processes following eruptions. Furthermore, we suggest that economical, political, and environmental pressures may make such interventions more common in the future. If volcano geoengineering ever becomes a discipline, it will need to

overcome many safety and ethical concerns, including dealing with uncertainty, deciding on philosophical approaches such as a consequentialism or precautionary principle, justice and inequality, military uses, cultural values, and communication. We highlight that while volcano geoengineering has significant potential benefits, the risks and uncertainties are too great to justify its use in the short term. Despite this, because of the potential large benefits to society, we believe there is a strong ethical case to support research into the efficacy and safety of volcano geoengineering for its potential future use. We propose that rigorous governance and regulation of any volcano geoengineering is required to protect against potential risks, to enable potentially valuable and publicly available research (e.g., quantification of efficacy and safety), to ensure that any future policy must be co-created through community engagement, and that volcano geoengineering should only be considered as part of larger mitigation practices.

Large language models and agricultural extension services a Perspective in *Nature Food*

Asaf Tzachor, Catherine Richards, CSER Affiliates with P. Pypers, A. Ghosh, J. Koo, S. Johal and B. King
6 November 2023

Several factors have traditionally hampered the effectiveness of agricultural extension services,

including limited institutional capacity and reach. Here we assess the potential of large language models (LLMs), specifically Generative Pre-trained Transformer (GPT), to transform agricultural extension. We focus on the ability of LLMs to simplify scientific knowledge and provide personalized, location-specific and data-driven agricultural recommendations. We emphasize shortcomings of this technology, informed by real-life testing of GPT to generate technical advice for Nigerian cassava farmers. To ensure a safe and responsible dissemination of LLM functionality across farming worldwide, we propose an idealized LLM design process with human experts in the loop.

Global history, the emergence of chaos and inducing sustainability in networks of socio-ecological systems in *PLOS ONE*

Sabin Roman from CSER with Francesco Bertolotti, CSER Visitor
17 November 2023

In this study, we propose a simplified model of a socio-environmental system that accounts for population, resources, and wealth, with a quadratic population contribution in the resource extraction term. Given its structure, an analytical treatment of attractors and bifurcations is possible. In particular, a Hopf bifurcation from a stable fixed point to a limit cycle emerges above a critical value of the extraction rate parameter. The stable

fixed-point attractor can be interpreted as a sustainable regime, and a large-amplitude limit cycle as an unsustainable regime. The model is generalized to multiple interacting systems, with chaotic dynamics emerging for small non-uniformities in the interaction matrix. In contrast to systems where a specific parameter choice or high dimensionality is necessary for chaos to emerge, chaotic dynamics here appears as a generic feature of the system. In addition, we show that diffusion can stabilize networks of sustainable and unsustainable societies, and thus, interconnection could be a way of increasing resilience in global networked systems. Overall, the multi-systems model provides a timescale of predictability (300-1000 years) for societal dynamics comparable to results from other studies, while indicating that the emergent dynamics of networks of interacting societies over longer time spans is likely chaotic and hence unpredictable.

Ineffective responses to unlikely outbreaks: Hypothesis building in newly-emerging infectious disease outbreaks in *Medical Anthropology Quarterly*

Freya Jephcott from CSER with James L N Wood, Andrew A Cunningham, J H Kofi Bonney, Stephen Nyarko-Ameyaw, Ursula Maier and P Wenzel Geissler
23 November 2023

Over the last 30 years, there has been significant investment in research and infrastructure aimed at mitigating the threat of newly emerging infectious diseases (NEID). Core epidemiological processes, such as outbreak investigations, however, have received little attention and have proceeded largely unchecked and unimproved. Using ethnographic material from an investigation into a cryptic encephalitis outbreak in the Brong-Ahafo Region of Ghana in 2010–2013, in this paper we trace processes of hypothesis building and their relationship to the organizational structures of the response. We demonstrate how commonly recurring features of NEID investigations produce selective pressures in hypothesis building that favor iterations of pre-existing “exciting” hypotheses and inhibit the pursuit of alternative hypotheses, regardless of relative likelihood. These findings contribute to the growing anthropological and science and technology studies (STS) literature on the epistemic communities that coalesce around suspected NEID outbreaks and highlight an urgent need for greater scrutiny of core epidemiological processes.

Technology Ties: the Rise and Roles of Military AI Strategic Partnerships an SSRN pre-print
Research Affiliate Matthijs Maas and Lena Trabucco (this work was produced during Lena’s visit to CSER)
29 November 2023

Emerging and Disruptive Technologies (EDTs) are reshaping military institutions and challenging traditional understandings of these actors’ tools, environment, and mission. This article explores how the proliferation of artificial intelligence (AI) technologies is set to drive a new generation of strategic partnerships, focused on nurturing and shaping that technology’s military application. These ‘military AI strategic partnerships’ are distinct from both traditional alliances and from strategic partnerships for other technologies, and may have significant implications for AI’s future military use, as well as for geopolitics. To investigate this phenomenon, this article explores the concept of a strategic technology partnership, discussing the uses, practices, and distinct operational requirements involved in military AI strategic partnerships. It then examines four recent cases: (1) the US-led AI Partnership for Defense (PfD); (2) the AUKUS partnership; (3) robust China-Russia cooperation on military AI; and (4) transatlantic (especially US-UK) cooperation. The article discusses the implications of these partnerships for broader AI governance stakeholders, standing military alliances like NATO, and for the development and usage of military AI itself.

3.2 Books

Recent CSER alumnus Lauren Holt produced a rich publication, blending commentary and creative writing, along with an audio artwork and illustrations that were produced as part of the project. The works are available on the CSER website.

Memetic Mythology for the End Times

Lauren Holt, CSER Affiliate
2 October 2023

Memetic Mythology for the End Times is a small book written by Dr Lauren Holt with the support of the V. Kann Rasmussen Foundation (VKRF). As part of efforts to address the current interlocking environmental and human social catastrophes VKRF term the 'poly-crisis', and with a focus on issues surrounding biodiversity, this text was written as a guide and support for using mythology and stories pre or post-civilizational collapse. Part performance art, part 'book of solace and sorcery' it is premised on the idea that should the worst happen there may be a way of rebuilding society on a better foundation than before through looking at stories that have been used in the past.



Lauren Holt's book Memetic Mythology for the End Times.



3.3 Reports

Lara Mani and Lalitha Sundaram co-produced an influential report with the United Nations Office for Disaster Risk Reduction (UNDRR) and other collaborators. Gideon Futerman and SJ Beard produced a comprehensive summary of a workshop on the relationships between solar radiation management technologies and catastrophic risk. CSER Affiliates were involved in two policy reports.

The Alan Turing Institute's response to the Large Language Models Inquiry: Call for Evidence
Fazl Barez, CSER Affiliate contributed with many others
5 September 2023

This document sets out The Alan Turing Institute's response to the House of Lords Communications and Digital Committee's Large Language Models Inquiry: Call for Evidence. The response synthesises the perspectives of researchers at the Turing with expertise and interest in the area of Large Language Models.

Hazards with escalation potential: Governing the drivers of global and existential catastrophes

Lara Mani, Lalitha Sundaram from CSER and Maxime Stauffer, Jenty Kirsch-Wood, Anne-Sophie Stevance, Sarah Dryhurst, Konrad Seifert
14 September 2023

A new report 'Hazards with Escalation Potential: Governing the Drivers of Global and Existential Catastrophes' was published by the UNDRR, International Science Council, CSER & Simon Institute for Longterm Governance.

The future of humanity and the planet hinges on human choices. How societies invest in critical infrastructure, political systems, military capacity and technological development creates both opportunities and risks. The impact of human activity has become so extensive that the risk of global and existential catastrophe is increasing fast.

What could cause global and existential catastrophe? What set of events and processes would lead to such worst-case scenarios? And what are the implications for risk research and governance?

This briefing note answers these questions by identifying the hazards that, once paired with corresponding vulnerabilities and exposures, would escalate and cause global and existential catastrophes. Its goal is to distil governance insights on risk cascades from a review of the literature, an expert survey and expert consultations.

Overall, out of the 302 hazards identified in the Hazard Information Profiles (HIPs) developed by the United Nations Office for

Disaster Risk Reduction (UNDRR) and the International Science Council to guide more holistic disaster risk reduction, 10 geological, biological, technological and social hazards were identified as having a global escalation potential. In addition to this list, climate change and artificial intelligence were identified as the most transformative processes with the potential to create, modify or amplify other hazards, vulnerabilities and exposures. This minority of known hazards, which could trigger cascades leading to global and existential catastrophe, warrants focus.

Escalating hazards share core characteristics such as the ability to affect multiple systems and to bypass established response and coping capacity. Focusing on these characteristics of the worst hazards can refine governance strategies, making them more adaptive to the various manifestations of risk. Current governance systems are built to prepare and respond to events with known frequency and manageable severity, but they are not fit for purpose to address worst-case scenarios, which are emerging, exponential and global in scope. This briefing note calls for important changes in risk research and governance to remedy these gaps.

[Open-Sourcing Highly Capable Foundation Models](#)
Elizabeth Seger, CSER Affiliate (and former CFI student fellow) with Noemi Dreksler, Richard

Moulange, Emily Dardaman, Jonas Schuett, K. Wei, et al
29 September 2023

Recent decisions by leading AI labs to either open-source their models or to restrict access to their models has sparked debate about whether, and how, increasingly capable AI models should be shared. Open-sourcing in AI typically refers to making model architecture and weights freely and publicly accessible for anyone to modify, study, build on, and use. This offers advantages such as enabling external oversight, accelerating progress, and decentralizing control over AI development and use. However, it also presents a growing potential for misuse and unintended consequences. This paper offers an examination of the risks and benefits of open-sourcing highly capable foundation models. While open-sourcing has historically provided substantial net benefits for most software and AI development processes, we argue that for some highly capable foundation models likely to be developed in the near future, open-sourcing may pose sufficiently extreme risks to outweigh the benefits. In such a case, highly capable foundation models should not be open-sourced, at least not initially. Alternative strategies, including non-open-source model sharing options, are explored. The paper concludes with recommendations for developers, standard-setting bodies,

and governments for establishing safe and responsible model sharing practices and preserving open-source benefits where safe.

Workshop Report: Managing the contribution of SRM and climate change to GCR

Gideon Futerman, CSER Visitor and SJ Beard
from CSER
21 November 2023

This report presents key findings from a workshop on managing the contribution of Solar Radiation Modification (SRM), a form of solar geoengineering, and Climate Change to Global Catastrophic Risk (GCR), which was hosted by Gideon Futerman and SJ Beard at the Centre for the Study of Existential Risk on March 28th and 29th 2023. The workshop was informed by a participatory futures exercise using the ParEvo technique that explored futures for SRM and SRM governance between 2030 and 2050, which some workshop participants took part in. Initial results of the exercise were shared with workshop participants and full results will be published separately.

Participants at the workshop emphasised that SRM can both contribute to and mitigate GCR; however, at present, high levels of uncertainty make it difficult to perform a complete assessment of risk. This report thus focuses on participants' exploration of the different pathways to global catastrophes and the role

SRM might play in them, the factors they saw as influencing the interaction between SRM and GCR, and their proposals for improving the governance of SRM and SRM research.

Participants identified many ways in which SRM may interact with GCR. Discussions of possible pathways towards global catastrophe typically involved interstate conflict, termination shock, and/or catastrophic climate impacts, while discussion of pathways away from global catastrophe involved the reduction of climate damages by SRM deployment. However, many other factors were also seen as influencing SRM's interaction with GCR. These included the type of deployment and governance, the perception of SRM's impacts and importance among politicians and publics, securitisation and militarisation, geopolitics, extreme weather, knowledge networks, wealthy individuals and corporations, and developments in artificial intelligence. Whether these interactions are net contributors or mitigators of GCR will depend on how they evolve and interact.

All these factors are contingent on human actions, perceptions, and behaviour. Ultimately, social, political, and geopolitical systems will be as important as physical systems in determining whether SRM reduces or increases GCR.

While it was generally felt that the current knowledge network around SRM has limited

influence, participants also believed that there were actions that could be taken to reliably reduce GCR and that this ought to be a consideration in research and policy and the report makes a number of recommendations based on these.

Safeguarding the safeguards: How best to promote AI alignment in the public interest

Seán Ó hÉigeartaigh from CSER with Oliver Guest and Michael Aird from the Institute for AI Policy and Strategy
14 December 2023

AI alignment work is important from both a commercial and a safety lens. With this paper, we aim to help actors who support alignment efforts to make these efforts as effective as possible, and to avoid potential adverse effects. We begin by suggesting that institutions that are trying to act in the public interest (such as governments) should aim to support specifically alignment work that reduces accident or misuse risks. We then describe four problems which might cause alignment efforts to be counterproductive, increasing large-scale AI risks. We suggest mitigations for each problem. Finally, we make a broader recommendation that institutions trying to act in the public interest should think systematically about how to make their alignment efforts as effective, and as likely to be beneficial, as possible.



UNIVERSITY OF
CAMBRIDGE

Contact

Amanda Lightstone
Head of Development — Arts and Humanities

Development and Alumni Relations
1 Quayside
Bridge Street
Cambridge, CB5 8AB

amanda.lightstone@admin.cam.ac.uk

Professor Matthew Connelly
Director – Centre for the Study of Existential Risk

16 Mill Lane
Cambridge CB2 1SB

director@cser.cam.ac.uk

(+44) 01223 760483

www.cser.ac.uk