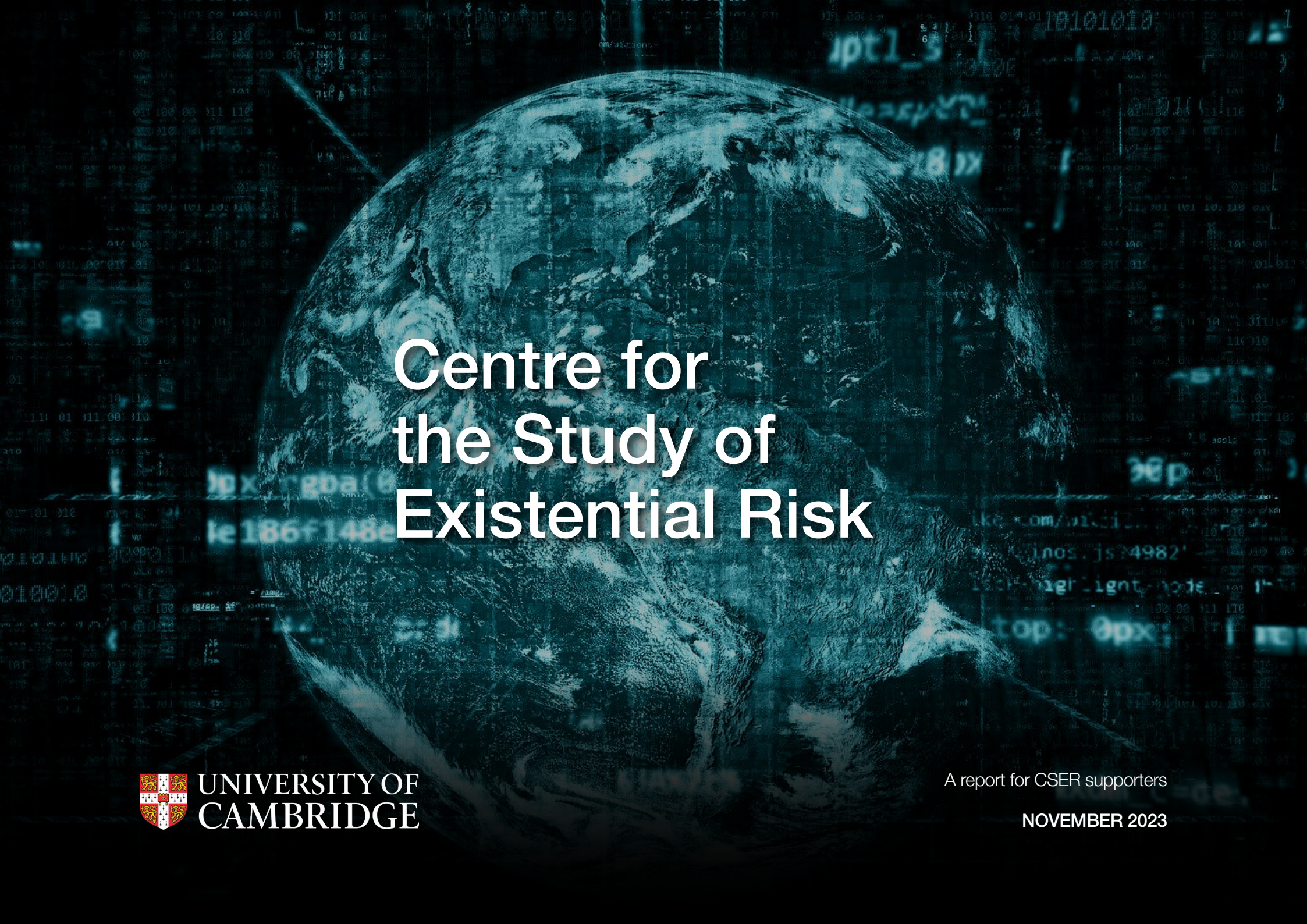


Centre for the Study of Existential Risk

UNIVERSITY OF CAMBRIDGE

A report for CSER supporters

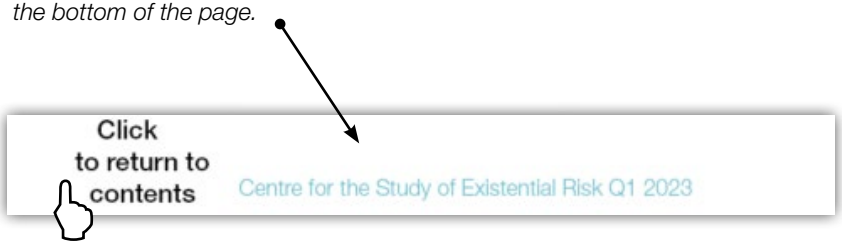
NOVEMBER 2023

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.



Contents

An introduction from **Matt Connelly**

Director, CSER

In July, I assumed the directorship of the Centre for the Study of Existential Risk. It is a dream job, since I have long been fascinated by the question of whether and how our species can overcome the many challenges we face – more and more of them challenges of our own making. And while I have been fortunate to have worked on several exciting research and policy problems over the last three decades, from nuclear proliferation to the politics of population control to AI risk-modelling, I have never before had the opportunity to lead such an impressive team with the breadth of expertise needed to take on the full range of catastrophic and existential risks. It is as if someone had blown the walls down that typically divide academic departments, assembled a task force of brilliant minds from multiple disciplines, and then put them to work on all the most dangerous threats confronting humanity.

It was both thrilling and sobering to take on these responsibilities. I spent much of July and August on a listening tour, trying to meet with all of CSER's many stakeholders. It was impossible to complete this task, since so many people have supported the Centre's important work. I learnt that many more look to us for leadership in this increasingly critical field.

I call it "critical" not just because it is so obviously important – more obvious with each new crisis – but also because there are so many

people with strong views on how we should define and prioritise the work to be done. But every person I spoke with agreed that CSER is at the very centre of it all, and that this is a uniquely important moment for leadership.

I am truly fortunate to be able to count on such excellent colleagues. In this report, I think you too will be amazed at all they have accomplished. This includes new discoveries published in top research journals, several secondments to important policymaking and advisory positions in government, and a host of public conversations and lectures designed for maximum impact, like that delivered by Geoff Hinton, one of the most significant events hosted this year at the University.

I am particularly grateful to the outgoing director, Seán Ó hÉigeartaigh, who continues to show incredible commitment to our cause. I also find myself depending on Jessica Bland, my deputy director, who displays remarkable insightfulness and savvy in helping set Centre strategy and managing the team from day-to-day. And all of us continue to be inspired by Lord Martin Rees, a co-founder of CSER. Improbably, he is not only one of the most important theoretical astrophysicists of his generation – recognised this past August by the Royal Society with the Copley Medal, one of the world's oldest and most prestigious prizes – he is also an astonishingly effective advocate for existential risk studies.

This report can only outline some of the work we have done in the last few months, and as it goes to press we are making even bigger plans for the year to come. But if you should have any thoughts or suggestions, please do be in touch. We are all in this together!

This report covers the period April to August 2023 and outlines our activities and future plans. Highlights of the last three months include:

- The conclusion of A Science of Global Risk, a project funded by Templeton World Charity Foundation, producing the book '*The Era of Global Risk*' edited by Martin Rees, SJ Beard, Catherine Richards and Clarissa Rios Rojas, with contributions from many other CSER staff and associates. The project also supported a number of workshops at CSER in early summer, marking the transition back to frequent in-person events.

- A public lecture and small seminar with Geoff Hinton hosted by CSER & CFI soon after his resignation from Google. The recording has 121,600 views on YouTube as of 31 August.

- The AI team started a period of intense engagement with governments, including a number of secondments that will continue into the autumn. Biological and natural risk researchers have been building new networks during the summer conference and seminar season, as well as taking meetings with influential stakeholders.

- This period saw an increase number of media commentaries including: an article in the Bulletin of *Atomic Sciences* from affiliate Kayla Lucero-Matteucci; articles in *Vox* and *The Conversation* from Haydn Belfied; and articles from affiliates Asaf Tzachor and Catherine Richards to accompany their paper on how to reduce Africa's undue exposure to climate risks.

- CSER and affiliates published **12 papers**, chapters and books. CSER researchers also produced **3 reports**, providing new advice and input into key public and policy discussions. Many of the papers included CSER affiliate and summer visitor Asaf Tzachor.

1. People

1.1 New Members of Staff

This term we have welcomed two new members to the CSER team.

Constantin Arnscheidt's research focuses on cascading global catastrophic risk: how might small stressors trigger system failure that leads to much larger catastrophes? He is attacking this problem using a diversity of approaches, but with a particular grounding in nonlinear dynamics, Earth science, and the study of complex systems. He has a PhD in Earth, Atmospheric, and Planetary Sciences from MIT, where he worked on the intersection of these topics, and an undergraduate degree in Physics from Harvard College.

Matthew Connelly became director of the Centre in July. Matthew is a professor of international and global history at Columbia University, and for the last seven years has been co-director of its social science research centre, the Institute for Social and Economic Research and Policy. Connelly comes with significant experience leading successful interdisciplinary initiatives focused on understanding and mitigating global catastrophic risk. From 2009-2013, Connelly directed the Hertog Global Strategy Initiative, a research program on the history and future of planetary threats, including nuclear war, pandemics, and climate change. Since then, Connelly has been the principal investigator of History Lab, a project that uses data science to analyse state secrecy, with a focus on intelligence, surveillance, and weapons of mass destruction. Connelly has taught courses on "The History and Future of Pandemic Threats and Global Public Health", "The History of the End of the World", and "The Future as History". He has frequently co-taught and co-authored articles with leading experts on pandemics, nuclear weapons, climate change, and religious violence.

Alice Jondorf is on a part-time secondment to job share with Clare Arnstein as Centre Coordinator, while Clare is on part-time secondment to CRASSH. She is combining this with her role as Centre Administrator at the Centre of Development Studies in the Department of POLIS.

Pat Wilson has joined CSER as Temporary PA/Office Manager. Her working life started in the Civil Service; she then moved to the commercial sector and for the past 15 years she has worked in various University Departments or Colleges in Cambridge. Pat read Classics at Newnham College, Cambridge.

1.2 Visiting Scholars

We have welcomed four new visitors during this period:

Kiana Tomita is a PhD student at the Graduate School of Advanced Integrated Studies in Human Survivability (GSAIS), University of Kyoto, Japan. She investigates what methods of communication are effective for disaster prevention at different phases of disasters under several climate change scenarios. Her research explores how to manage future disasters by improving communication, increasing risk awareness, and educating people about disasters. Kiana studies how to urge communities to evacuate during floods, using Japan as a case study. She also holds an MPhil in East Asia and Middle Eastern Studies from the University of Cambridge. She was a visitor from April to October 2023.

Sarah Woods is an award-winning playwright and Associate Professor at the Denmark National School of Performing Arts. She will be working with Paul Ingram on their project People & Patterns: transforming the ways we think and connect when everything is at risk. She is a visitor from April 2023 to January 2024.

Alexander Saeri works to increase the reach and impact of behaviour science on the world's most pressing problems. He uses a mix of applied behaviour science and social science methods to understand and address complex challenges, including climate change, pandemics, and artificial intelligence. He is especially interested in identifying "WHO needs to do WHAT differently" in the context of global catastrophic and existential risk communication, developing and evaluating mixed method interventions to influence decision making and behaviour, and scaling up effective interventions. He was a visitor at CSER in May and June 2023.

Benoit Pelopidas is the founding director of the Nuclear Knowledges program (formerly chair of excellence in security studies) at Sciences Po (CERI). His program, "Nuclear Knowledges", is the first independent scholarly research program on the nuclear phenomenon in France. He is also an affiliate of the Center for International Security and Cooperation (CISAC) at Stanford University. His research has received four international prizes and the most prestigious European grants based on scholarly assessment by peers, most notably an ERC Starting Grant. This interdisciplinary effort of independent scholarship has led to the following discoveries over the last five years: the lack of credibility and rationality of the French nuclear arsenal at least until 1974; the underestimation of the effects of French nuclear weapons tests in Polynesia; the role of luck in the past avoidance of unwanted nuclear explosions; the limits of popular support for nuclear weapons policy, the role of nostalgia and imagined futures in shaping nuclear weapons politics and the effects of funding carrying conflicts of interests on nuclear weapons policy analysis. He was a visitor in June 2023, when he also gave a public lecture.

Asaf Tzachor worked at CSER until 2021 and is a current research affiliate. Asaf is an interdisciplinary researcher, practitioner, and educator at the interface of sustainability sciences, emerging technologies, and global risks. He is an Associate Professor for Sustainability at Reichman University. He visited CSER in summer 2023 to reconnect with CSER researchers and help develop a research agenda.

1.3 Research Affiliates

We have welcomed one new research affiliate, who also joins CSER's advisory board.

Madhulika Srikumar is the Program and Research Lead at Partnership on AI (PAI), overseeing the Safety-Critical AI program. PAI is a global non-profit partnership of leading industry, academic, and civil society organisations, advancing best practices in AI governance. Madhulika leads a team that develops best practices through participatory processes to provide actionable guidance that can be adopted in practice by PAI's Partners, inform public policy, and advance public understanding. At CSER, she collaborates with the AI: FAR team to explore geopolitical implications of proposals for regulating compute to mitigate existential risk – in particular, whether there is a tension to be addressed between equity and safety when determining how the majority world gets access to large compute.

1.4 Leavers

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

- Anna Chau left to start a PhD at the UCL Institute for Risk and Disaster Reduction, working on warning systems and decision making.

- Lauren Holt published 'Memetic Mythology For the End Times', a booklet based on her work at CSER and has recently signed a book deal with Penguin Random House.

- Ellen Quigley now co-directs the Finance for Systemic Change Centre in the Department of Land Economy. On 3 July, CSER held a meeting to celebrate her team's excellent work and welcomed back alumni from the Sustainable Finance team, including Mia Sannapureddy, Akaraseth Puranasamriddhi and Jake Ainscough.

2. Events, Engagement and Outreach

2.1 Academic engagement

Researchers mainly focused on meetings in their specialist areas during the summer season of academic conferences. This included Freya Jephcott's engagement with infectious disease specialists, and a summer series of her Hidden Epidemics seminars. Lara Mani presented to a high profile gathering of geoscientists and an interdisciplinary workshop with historians.

- 4-5 April: Freya Jephcott attended the Cambridge Infectious Diseases Annual Symposium
- 24-28 April: Lara Mani delivered an invited keynote and another talk at the European Geoscience Union, Vienna
- 3 May: Freya Jephcott met with Associate Prof Seye Abimbola, Editor of BMJ Global Health
- 4-5 May: Freya Jephcott gave a seminar at the University of Sydney titled Ineffective Responses to Unlikely Outbreak and met with Professor Jaime Miranda, Head of the School of Global Health at the University of Sydney
- 10-14 May: SJ Beard co-organized a workshop at Wytham Abbey on Pluralisms in Existential Risk Studies
- 22-24 May: Lara Mani gave a keynote talk at a workshop in Bern Switzerland held by the Volcanic Impacts on Climate and Society (VICS) Working Group, part of the Past Global Changes (PAGES) program within Future Earth
- 25 May: Lara Mani presented a seminar at the University of Geneva on the global risks posed by volcanic eruptions
- June: Lalitha Sundaram and affiliate Charlotte Hammer attended a series of discussions with Riesgos Catastroficos Globales, the Spanish language network for studying Global Catastrophic Risks to refine their biosecurity programme
- 2 June: Lalitha Sundaram gave a talk at the University of Oxford as part of the Grand Challenge Seminar: From deepfakes to deadly viruses: Governance and ethics in science
- 16 June: Freya Jephcott spoke on a panel on Interspecies and Multispecies Epizootics at the University of St Andrews' conference Epizootics Beyond the Farm
- 23 June: Jochem Rietveld attended the Herrenhausen Conference on Climate crisis and systemic risks: Lessons Learned, developing potential collaborations for CSER's COVID Lessons project
- 27 June 2023: Lalitha Sundaram attended the Engineering Biology Interdisciplinary Research Centre Steering Group Retreat

- 25 July and 1 August: Freya Jephcott hosted a summer series of Hidden Epidemics seminars with talks from with Coreen McGuire, University of Edinburgh and John Aggrey, Virginia Tech

- 29 July: Haydn Belfield presented a paper 'Compute and Antitrust: Regulatory implications of the AI hardware supply chain, from chip design to foundation model APIs' at a workshop on Generative AI and Law (GenLaw '23) at ICML 20238-10

- 8-10 August: Haydn Belfield and affiliate Shin-Shin Hua presented a paper 'Effective Enforceability of EU Competition Law Under AI Development Scenarios: a Framework for Anticipatory Governance' at the Sixth Annual AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society in Montreal

- 9 August: CSER launched its methodological toolkit, an accessible introduction to some of the methods and tools that we have developed and used within CSER to encourage and support others in engaging with them in their own work

- 16 August: Several CSER researchers signed a statement on Pluralism in Existential Risk Studies that was developed by CSER visitor Gideon Futerman following a workshop that they helped organise in May 2023.

2.2 Policy Engagement

CSER researchers have been exploring new policy connections, with a long list of meetings including a visit from the Office of the Director of National Intelligence in the US and the new UK Cabinet Office's Resilience Directorate. Lalitha Sundaram has taken a position on the UK government's newly-formed Biosecurity Leadership Council. The AI team started a period of intense engagement with governments, which extends to secondments that will continue into the autumn. CSER visitor Taniel Yusef took our agenda to many stakeholders in international security and diplomacy.

- 2-7 April: Lara Mani spoke as part of a panel on disaster management and the role of international agencies for NEO impacts at the 8th IAA Planetary Defense Conference at the UN in Vienna: Watch from 01:01:24

- 14 April: Lalitha Sundaram met with Logan Graham from Anthropic

- 26 April: Lalitha Sundaram had a meeting with the Council on Strategic Risks

- 27 April: The Deputy Director of the Strategic Futures Group at the Office of the Director of National Intelligence in the US visited CSER. Their team produces the Global Trends report ahead of each US Presidential term

- 28 April: Tom Hobson, Alex Klein and Lalitha Sundaram attended a workshop with civil society and policy leaders on Learning from the Past and Looking to the Future after Review Conference: Integrating NGO work on Codes of Conduct and an International Biological Security Education Network into the next BTWC Intersessional Process, organised by the Biological Security Research Centre at London Metropolitan University

- 5 May: Freya Jephcott met Dr Sarah Hill, from New Zealand's Royal Commission into Covid

- 10 May: Shahar Avin and Haydn Belfield participated in the UK Government Office of Science's Trajectories towards Artificial General Intelligence Workshop and continue to work with them on their AI scenarios

- 11 May: Lara Mani met with Rashmin Gunasekera, GFDRR, World Bank

- 11 May: Lalitha Sundaram met with Ian Hogarth and Carina Namih, Plural Platform

- 16 May: Lalitha Sundaram presented at the Systemic Risks Crash Course for Asset Owners, organised by Ellen Quigley and the Sustainable Finance team

- 17 May: Freya Jephcott met Nancy Griffiths from the US Department of Defence about the design of pandemic early warning systems

- 17 May: CSER visitor Taniel Yusef spoke at the side event "Perspectives on Unpredictability in Autonomous Weapons Technology" at the Group of Governmental Experts on Lethal Autonomous Weapons Systems

- 17 May: Jess Bland spoke at Policy Horizons Canada Futures Week on a panel to launch their new briefing paper on global existential risks

- 18 May: CSER visitor Taniel Yusef organised and spoke on a panel about risk mitigation and confidence measures at the Group of Governmental Experts on Lethal Autonomous Weapons Systems.

- 19 May: A paper Lalitha Sundaram had written on synthetic biology regulation was circulated among the Engineering Biology Leadership Council, as they discussed regulation detailed here

- 19 May: Shahar Avin and Haydn Belfield presented on AI risks as part of the UK Foreign, Commonwealth & Development Office Geopolitics Directorate away day and Shahar has subsequently been added to an advisory group

- 23 May: The CSER team were visited by officials from the Cabinet Office's National Security Secretariat and Resilience Directorate to learn more about our work and advance government's thinking on existential risks as part of developing a relationship with CSER over the coming months

- 23 May: CSER visitor Taniel Yusef spoke on the panel "Exploring the Impact of Autonomous Weapons Usage: A NextGen Discussion" held by the Royal United Services Institute

- 25 May: Lara Mani met with Ian Lisk of WMO/Met Office in Geneva to discuss a potential secondment to the WMO for volcanic risk.

- 25 May: Shahar Avin and affiliate Kayla Lucero-Matteucci participated in a Defense Acquisition University and Centre for Data Ethics and Innovation AI ethics case study with EC2 Spearhead

- 27-28 May: Shahar Avin organised a team retreat for Technology Strategy Roleplay (a charity that spun out of work at CSER developing the game Intelligence Rising) where they created new designs for Intelligence Rising

- 7 June: Freya Jephcott attended a Data & Disease workshop run by the University of Edinburgh

- 15 June: Freya Jephcott, Lalitha Sundaram, and affiliate Luke Kemp attended the OECD Strategic Foresight Unit's workshop on biotechnology futures in Oxford

- 16 June: Freya Jephcott met with Nancy Griffiths from the UK Department of Defence about designing pandemic early warning systems

- 19 June: Lara Mani met with Henry Green, Head of Net Zero Systems and Delivery in the Department for Energy Security and Net Zero (DESNZ)

- 20 June: Lalitha Sundaram attended the CSaP Policy Workshop on Engineering Biology

- 20 June: Freya Jephcott attended Epistemic Justice for Healthy Societies held by the Sydney Centre for Healthy Societies

- 26 June: Jess Bland met Lord Toby Harris and Katie Barnes from the National Preparedness Commission to discuss future collaborations

- 26 June: Tom Hobson took part in a workshop hosted by the University of Bath focused on the UK Chemical and Biological Weapons incident preparedness

- 29 June: Lalitha Sundaram, Luke Kemp and Lara Mani contributed to the ASRA (Accelerator for Systemic Risk Assessment) Online consultation

- 30 June: Lara Mani met with Simon Baugh, Chief Executive of the Government Communications Service in the Cabinet Office to discuss risk communication work at CSER.

- 5 July: Shahar Avin and CFI fellow Charlotte Stix organised a workshop on Evaluation of AI models for dangerous capabilities

- 11 July: Lalitha Sundaram attended a roundtable organised by the Secretary of State for Science, Innovation and Technology on “Public interest in, and uptake of, engineering biology”

- 17-19 July: Haydn Belfield attended an ELN Workshop on AI and nuclear weapons in July. Alice Saltini discusses some of the issues here in a post afterwards

- 18 July: Jess Bland attended a workshop on Space Policy narratives story listening hosted by colleagues in Cambridge, Oxford and government partners.

2.3 Public Engagement

This period saw an increase number of media commentaries including: an article in the *Bulletin of Atomic Sciences* from affiliate Kayla Lucero-Matteucci; articles in *Vox* and *The Conversation* from Haydn Belfied; and articles from affiliates Asaf Tzachor and Catherine Richards to accompany their paper on how to reduce Africa’s undue exposure to climate risks.

- 4 April: Haydn Belfield published an article in the Conversation on how the EU and US are steaming ahead of the UK on AI regulation and standards

- 13 April: TERRA was re-released after a period of downtime due to maintenance

- 21 April: Haydn Belfield spoke at EAGlobal Nordics about how the Nordic countries could be world-leaders in reducing existential risk

- 1 May: CSER Affiliate Kayla Lucero-Matteucci wrote an article in the Bulletin of the Atomic Scientists on how catastrophic risks are converging, arguing that researchers need to step out of their silos

- 15 May: Continued media interest in the paper "Global catastrophic risk from lower magnitude volcanic eruptions" by Lara Mani, Asaf Tzachor and Paul Cole included TikTok videos about the paper

- 19 May: Paul Ingram spoke on a panel exploring the future of arms control, managing nuclear tensions with Russia and China's changing nuclear strategy

- 19 May: Haydn Belfield interviewed Martin Rees during an informal discussion at the EAGx in Cambridge about issues related to astronomy, space travel and catastrophic risk.

- 19 May: Lara Mani presented a lightning talk about volcanic risk (from 12 min) and Paul Ingram spoke on a panel on nuclear risk reduction at EA Global London

- 31 May: Seán Ó hÉigeartaigh spoke on a panel titled "Algorithms Against Humanity" and appeared on the In Reality podcast with Eric Schurenberg titled "The podcast about truth, disinformation and the media" at the Dublin Tech Summit

- 31 May: Martin Rees spoke to CBC Radio about his recent book '*If Science is to Save Us*' and how science needs to become part of our common culture

- 6 June: Martin Rees, Paul Ingram and Lara Mani spoke on the Templeton World Foundation Stories of Impact podcast about CSER's research

- 18 June: CSER Visitor Ben Holt wrote a story for the Association of Professional Futurists about a time travel visit to 2450

- 30 June: SJ Beard appeared on BBC Radio 3's Free Thinking to discuss Dystopias

- 19 June: Paul Ingram appeared on TRT several times to discuss if Russian nuclear warheads in Belarus are a threat and NATOs possible involvement in Ukraine

- 11 July: Seán Ó hÉigeartaigh appeared on the New Thinking podcast to discuss what generative AI means for the information landscape

- 22 July: Haydn Belfield wrote an article for Vox about Robert Oppenheimer and the ways Christopher Nolan's film gets the famous scientist wrong

- 14 August: Authors including CSER research affiliates Asaf Tzachor and Catherine Richards wrote an article for *Nature* titled "How to reduce Africa's undue exposure to climate risks" Subsequent media coverage included articles in AP News, La Vanguardia and on the University of Cambridge website. On 21 August, Asaf and Catherine wrote a follow up article in The Conversation

- 31 August 2023: The *Cambridge Independent* wrote about the Existential Risk Alliance (ERA) Cambridge Fellowship led by Nandini Shiralkar who spoke at the CSER CCCR 2022 conference and several CSER researchers had the pleasure of mentoring fellows this year.

2.4 Events

The beginning of this period included a succession of workshops, marking the final transition back to in-person events. In May, CSER and CFI hosted a public lecture from Geoff Hinton, which as of 31 August has 121,600 views on YouTube.

- 4 April: Alex McLaughlin hosted a workshop on climate protest and resistance. As well as foregrounding philosophical

perspectives on the issue, the workshop drew from empirical work that aims to understand the prospects of success for different forms of action

- 12 April: CSER hosted a visit and internal talk by Professor Christopher Chyba, Professor of Astrophysical Sciences and International Affairs at Princeton University. Professor Chyba spoke on New Technologies and Nuclear Escalation.

- 13 April: Tom Hobson, Lalitha Sundaram and Alex Klein hosted the Ninth Review Conference of the Biological Weapons Convention: Where Next? workshop. The workshop brought together experts from UK civil society to discuss developments in biological security governance in the wake of the Ninth RevCon

- 24 April: Paul Ingram and CSER visitor Sarah Woods hosted three workshops as part of their People & Patterns project. The workshop to explore nuclear cultures on 24 April focused on deepening our capacity to engage with diverse nuclear cultures. On 11 May, the creating spaces workshop aimed to explore how creative systems methodologies can help us to engage with the complex world around us and the final workshop "Inter-disciplinary Exploration Into How We Think About Global Risk" held on 20 June, explored tools that take a whole systems approach to the global risks we face. On 11 July, Paul and Sarah ran a workshop to introduce their People and Patterns project at the International Institute for Applied Systems Analysis

- 25 May: CSER and CFI hosted a public lecture from Geoff Hinton titled "Two Paths to Intelligence". The event, pictured right, sold out and as of 31 August has had 121,600 views on YouTube

- 14 June: Professor Benoît Pelopidas held the public lecture "Scoping Nuclear Weapons Choices in an Age of Existential Threats"

- 19 July: CSER, the Centre for Geopolitics and hosted the panel "Nuclear Risk Reduction in the Baltic Sea Region" with support from the European Leadership Network. The panel included Artis Pabriks, Dr Marion Messmer, Rt Hon Charles Clarke and CSER's Paul Ingram.

3. Publications

3.1 Papers

Lalitha Sundaram wrote a paper with Cambridge colleagues discussing how historical concepts based on containment and release frame the regulation of Synthetic Biology.

Synthetic Biology Regulation in Europe: Containment, Release, and Beyond in *Synthetic Biology* on 10 May by Lalitha S Sundaram, James W Ajiokan and Jennifer C Molloy

While synthetic biology is hoped to hold promise and potential to address pressing global challenges, the issue of regulation is an under-appreciated challenge. Particularly in Europe, the regulatory frameworks involved are rooted in historical concepts based on containment and release. Through a series of case studies including a field-use biosensor intended to detect arsenic in well water in Nepal and Bangladesh, and insects engineered for sterility, we explore the implications that this regulatory and conceptual divide has had on the deployment of synthetic biology projects in different national contexts. We then consider some of the broader impacts that regulation can have on the development of synthetic biology as a field, not only in Europe but also globally, with a particular emphasis on low- and middle-income countries. We propose that future regulatory adaptability would be increased by moving away from a containment and release dichotomy and toward a more comprehensive assessment that accounts for the possibility of varying degrees of 'contained release'.

Shahar Avin joined CSER Affiliates Catherine Richards and Asaf Tzachor in co-authoring a paper that recommends interventions for safe and responsible deployment of AI across water supply and sewage systems, including prioritising applications based on their benefit.

Rewards, risks and responsible deployment of artificial intelligence in water systems in *Nature Water* on 11 May by Catherine Richards, Asaf Tzachor, Shahar Avin and Richard Fenner.

Artificial intelligence (AI) is increasingly proposed to address deficiencies across water systems, which currently leave about 25% of the global population without clean water, about 50% without sanitation services and about 30% without hygiene facilities. AI is poised to enhance supply insights, catchment management and emergency response, improve treatment plant and distribution network design, operation and maintenance, and advance service availability, demand management and water justice. However, proliferation of this nascent technology could trigger serious and unexpected problems, including system-wide compromise owing to design errors, malfunction and cyberattacks as well as exposures to cascading socio-ecological, water–energy–food nexus and coupled critical infrastructure failures. In response, we make three recommendations for safe and responsible deployment of AI across potable water supply and sewage disposal systems: address gaps in foundational infrastructure and digital literacy; establish institutional, software

and hardware mechanisms for trustworthy AI; and prioritize applications based on our proposed systematic benefit and risk assessment framework.

Affiliate Catherine Richards co-authored a paper on how runaway global warming could affect food systems.

International risk of food insecurity and mass mortality in a runaway global warming scenario in *Futures* on 31 May by Catherine Richards, Hannes Gauch and Julian Allwood.

Climate and agriculture have played an interconnected role in the rise and fall of historical civilizations. Our modern food system, based on open-environment production and globalised supply chains, is vulnerable to a litany of abiotic and biotic stressors exacerbated by anthropogenic climate change. Despite this evidence, greenhouse gas emissions continue to rise. Current trajectories suggest global warming of ~2.0–4.9 °C by 2100, however, a worst-case emissions scenario with rapid combustion of all available fossil fuels could cause a rise of ~12 °C. Even if emissions decline, unprecedented atmospheric CO2-e concentrations risk triggering tipping points in climate system feedbacks that may see global warming exceed 8 °C. Yet, such speculative 'runaway global warming' has received minimal attention compared to mainstream low- to mid-range scenarios. This study builds on The Limits to Growth to provide new insights into the international risk of mass mortality due to food insecurity based on a higher-resolution illustration of World3's 'runaway global warming' scenario (~8–12 °C+). Our simulation indicates rapid decline in food production and unequal distribution of ~6 billion deaths due to starvation by 2100. We highlight the importance of including high-resolution simulations of high-range global warming in climate change impact modelling to make well-informed decisions about climate change mitigation, resilience and adaptation.

Alex Marcoci co-authored a study led by former CSER affiliate Bonnie Wintle looking at how replicable results were in structured group exercises to elicit information. They showed there is some evidence that exercises that engaged in a greater breadth of reasoning or with greater statistical literacy among the group provided more accuracy.

Predicting and reasoning about replicability using structured groups in *Royal Society Open Science* on 7 June by Bonnie Wintle, Eden T. Smith, Martin Bush, Fallon Mody, David P. Wilkinson, Anca M. Hanea, Alex Marcoci, Hannah Fraser, Victoria Hemming, Felix Singleton Thorn, Marissa F. McBride, Elliot Gould, Andrew Head, Daniel G. Hamilton, Steven Kambouris, Libby Rumpff, Rink Hoekstra, Mark A. Burgman and Fiona Fidler

This paper explores judgements about the replicability of social and behavioural sciences research and what drives those judgements. Using a mixed methods approach, it draws on qualitative and quantitative data elicited from groups using a structured approach called the IDEA protocol ('investigate', 'discuss', 'estimate' and 'aggregate'). Five groups of five people with relevant domain expertise evaluated 25 research claims that were subject to at least one replication study. Participants assessed the probability that each of the 25 research claims would replicate (i.e. that a replication study would find a statistically significant result in the same direction as the original study) and described the reasoning behind those judgements. We quantitatively analysed possible correlates of predictive accuracy, including self-rated expertise and updating of judgements after feedback and discussion. We qualitatively analysed the reasoning data to explore the cues, heuristics and patterns of reasoning used by participants. Participants achieved 84% classification accuracy in predicting replicability. Those who engaged in a greater breadth of reasoning provided more accurate replicability judgements. Some reasons were more commonly invoked by

more accurate participants, such as ‘effect size’ and ‘reputation’ (e.g. of the field of research). There was also some evidence of a relationship between statistical literacy and accuracy.

John Burden, Seán Ó hÉigeartaigh and other colleagues from across Cambridge and beyond wrote a paper on a “human-centred generality” (HCG), rather than a fully autonomous general intelligence.

Your Prompt is My Command: On Assessing the Human-Centred Generality of Multimodal Models in *Journal of Artificial Intelligence Research* on 12 June by Wout Schellaert, Fernando Martınez-Plumed Karina Vold John Burden Pablo A. M. Casares Roi Reichart Sean O hEigeartaigh Anna Korhonen and Jose Hernandez-Orallo

Even with obvious deficiencies, large prompt-commanded multimodal models are proving to be flexible cognitive tools representing an unprecedented generality. But the directness, diversity, and degree of user interaction create a distinctive “human-centred generality” (HCG), rather than a fully autonomous one. HCG implies that —for a specific user— a system is only as general as it is effective for the user’s relevant tasks and their prevalent ways of prompting. A human-centred evaluation of general-purpose AI systems therefore needs to reflect the personal nature of interaction, tasks and cognition. We argue that the best way to understand these systems is as highly-coupled cognitive extenders, and to analyse the bidirectional cognitive adaptations between them and humans. In this paper, we give a formulation of HCG, as well as a high-level overview of the elements and trade-offs involved in the prompting process. We end the paper by outlining some essential research questions and suggestions for improving evaluation practices, which we envision as characteristic for the evaluation of general artificial intelligence in the future. This paper appears in the AI & Society track.

John Burden joined a group of authors including CSER affiliates David Krueger and frequent collaborator Adrian Weller to highlight harms from algorithmic systems that are not entirely under human control and ways forward for addressing them.

Harms from Increasingly Agentic Algorithmic Systems in *FAccT ‘23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* on 12 June by Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi,Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger and Tegan Maharaj

Research in Fairness, Accountability, Transparency, and Ethics (FATE) has established many sources and forms of algorithmic harm, in domains as diverse as health care, finance, policing, and recommendations. Much work remains to be done to mitigate the serious harms of these systems, particularly those disproportionately affecting marginalized communities. Despite these ongoing harms, new systems are being developed and deployed which threaten the perpetuation of the same harms and the creation of novel ones. In response, the FATE community has emphasized the importance of anticipating harms. Our work focuses on the anticipation of harms from increasingly agentic systems. Rather than providing a definition of agency as a binary property, we identify 4 key characteristics which, particularly in combination, tend to increase the agency of a given algorithmic system: underspecification, directness of impact, goal-directedness, and long-term planning. We also discuss important harms which arise from increasing agency – notably, these include systemic and/or long-range impacts, often on marginalized stakeholders. We emphasize that recognizing agency of algorithmic systems does not absolve or shift the

human responsibility for algorithmic harms. Rather, we use the term agency to highlight the increasingly evident fact that ML systems are not fully under human control. Our work explores increasingly agentic algorithmic systems in three parts. First, we explain the notion of an increase in agency for algorithmic systems in the context of diverse perspectives on agency across disciplines. Second, we argue for the need to anticipate harms from increasingly agentic systems. Third, we discuss important harms from increasingly agentic systems and ways forward for addressing them. We conclude by reflecting on implications of our work for anticipating algorithmic harms from emerging systems.

Affiliates Catherine Richards, Tom Cernev and Asaf Tzachor published a paper calling for urgent expansion of existing "responsible use of AI in space"

Safely advancing a spacefaring humanity with artificial intelligence published in *Frontiers* on 15 June by Catherine Richards, Tom Cernev, Asaf Tzachor, Gustavs Zilgalvis and Bartu Kaleagasi

A "Space Renaissance" is underway. As our efforts to understand, utilize and settle space rapidly take new form, three distinct human-space interfaces are emerging, defined here as the "Earth-for-space," "space-for-Earth" and "space-for-space" economies. Each engenders unprecedented opportunities, and artificial intelligence (AI) will play an essential role in facilitating innovative, accurate and responsive endeavors given the hostile, expansive and uncertain nature of extraterrestrial environments. However, the proliferation of, and reliance on, AI in this context is poised to aggravate existing threats and give rise to new risks, which are largely underappreciated, especially given the potential for great power competition and arms-race-type dynamics. Here, we examine possible beneficial applications of AI through the systematic prism of the three economies, including advancing the astronomical sciences, resource efficiency, technological innovation, telecommunications, Earth observation, planetary defense, mission strategy, human life support systems and artificial astronauts. Then we consider unintended and malicious risks arising from AI in space, which could have catastrophic consequences for life on Earth, space stations and space settlements. As a response to mitigate these risks, we call for urgent expansion of existing "responsible use of AI in space" frameworks to address "ethical limits" in both civilian and non-civilian space economy ventures, alongside national, bilateral and international cooperation to enforce mechanisms for robust, explainable, secure, accountable, fair and societally beneficial AI in space.

Lara Mani and her collaborators from The University of the West Indies Seismic Research Centre, published an evaluation of their work on the crisis communications during a volcanic eruption.

Evaluating the crisis communications campaign during the 2020-2021 eruption of La Soufrière, St Vincent in a Special Publication for the Geological Society in July 2023 by Lara Mani, Stacey Edwards, Erouscilla Joseph, Alia Juman and Thalia Thomas

During the 2020–21 eruption of La Soufrière, St Vincent, the University of the West Indies, Seismic Research Centre played a major role in supporting communication of hazard and risk information to publics and stakeholders across St Vincent. Due to COVID-19 restrictions on in-person education and outreach activities, the communications campaign was heavily reliant on social media platforms, and TV and radio broadcasts. Although the communications approach sought to be inclusive of all members of the affected communities, we consider that more vulnerable residents, such as the elderly, children, and those with low literacy levels and limited digital access were likely excluded from the communication efforts.

In order to establish effectiveness of the crisis communications campaign at engaging communities and stakeholders with relevant information, and to identify areas for improvement, a large-scale evaluation campaign was conducted in St Vincent in August 2021. The results demonstrate that radio broadcasts are the most important communication tool for broad community reach, but that person-to-person information sharing was more important in the most exposed communities. Agencies such as the Red Cross and grassroots community disaster preparedness groups were instrumental in amplifying the reach of information to vulnerable members of at-risk communities and for evacuation co-ordination.

Tom Hobson, Lara Mani, affiliate Catherine Rhodes and Lalitha Sundaram published a chapter in a new book reflecting on the recent pandemic and its ramifications.

Chapter 11: Evaluating COVID-19 in the Context of Global Catastrophic Risk in *Evaluating a Pandemic* in August 2023 by Tom Hobson, Lara Mani, Catherine Rhodes, and Lalitha Sundaram

Overall, whether or not COVID-19 fits with particular definitions of global catastrophic risk (GCR), it provides a case from which researchers, policy makers and practitioners can learn and improve their understanding of how GCRs and responses to them might play out. Likewise, scholarship from the field of existential and GCR studies, and from global catastrophic biological risk (GCBR) studies in particular, can help inform broader understanding of the pandemic.

Alex Marcoci and others highlighted the widening gap between the science, technology, engineering and mathematics disciplines and the humanities and social and behavioural sciences, and how this is damaging our ability to see the bigger picture and draw important insights from STEM work.

Big STEM collaborations should include humanities and social science in *Nature Human Behaviour* on 14 August by Alex Marcoci Ann C. Thresher, Niels C. M. Martens, Peter Galison, Sheperd S. Doeleman and Michael D. Johnson

The divide between the natural sciences and the humanities and social sciences in the West is a recent one. Newton considered himself a 'natural philosopher', Thomas Hobbes thought that one of his greatest achievements was laying the foundations of optics, and Margaret Cavendish was the author of one of the first works of science fiction and the first woman to attend a meeting of the Royal Society. More recently, the space between the so-called STEM ('science, technology, engineering and mathematics') disciplines and the humanities and social and behavioural sciences has widened, until we have come to see them as islands without bridges.

Affiliates Asaf Tzachor and Catherine Richards published an analysis with climate science colleagues in Kenya and Senegal on the increased risk from climate change on the African continent.

How to reduce Africa's undue exposure to climate risks in *Nature* on 14 August by Asaf Tzachor, Catherine Richards, Masilin Gudoshava, Patricia Nying'uro, Herbert Misiani, Jemimah G. Ongoma, Yoav Yair, Yacob Mulugetta and Amadou T. Gaye

Climate and weather-related disasters, including tropical cyclones, storm surges, floods and droughts, are on the rise. Over the past 50 years, rates have increased fivefold globally, and the damages associated with them have swelled by 70 times. This will only get worse as climate change increases the frequency and intensity of extreme weather. And some places are feeling the brunt much more than others — notably Africa.

The output of a CSER project called A Science of Global Risk, the book '*The Era of Global Risk*' was edited by Martin Rees, SJ Beard, Catherine Richards and Clarissa Rios Rojas, with contributions from many other CSER staff and associates. The book is published open access and the digital version is freely available from the publisher. SJ Beard wrote a blog to accompany the launch.

The Era of Global Risk: An Introduction to Existential Risk Studies on 31 August by Martin Rees, S. J. Beard, Catherine Richards, Clarissa Rios Rojas, Rachel Bronson, Sabin Roman, Lalitha Sundaram, Natalie Jones, Sheri Wells-Jensen, Lara Mani, Doug Erwin, Lindley Johnson, Luke Kemp, Kobi Leins, Nancy Connell, John Burden, Sam Clarke, Jess Whittlestone, Matthijs Maas, Kayla Lucero-Matteucci and Di Cooke.

This volume presents a series of specially written essays that explore different aspects of global risk, with the potential to bring about human extinction and civilization collapse. Bringing together experts from many disciplines working at or collaborating with CSER, it provides a comprehensive survey of what we know about this risk, how we can understand it better, and, most importantly, what can be done to manage it effectively.

These essays pair insights from decades of research and activism around global risk with the latest academic findings from the emerging transdisciplinary field of Existential Risk Studies. They assess natural systems, societal pressures, and technological advances to build an empowering vision of how we can safeguard humanity's long-term future.

The book covers both methods and approaches for studying and managing global risk with in-depth discussion of core risk drivers: including environmental breakdown, novel technologies, global scale natural disasters, and security threats. It is aimed to be both Inspiring and accessible for students of global risk and those committed to its mitigation and poses the critical question: how can we make sense of this era of global risk and move beyond it to an era of global safety?

3.2 Reports

Maurice Chiodo co-authored a manifesto for responsible development of mathematical works as a practical tool and aid for anyone carrying out, managing or influencing mathematical work. the product of seven years of work, it provides insight into how to undertake and develop mathematically-powered products and services in a safe and responsible way. The authors intend to make revisions to this document over time.

Manifesto for the Responsible Development of Mathematical Works – A Tool for Practitioners and for Management on arXiv on 15 June by Maurice Chiodo and Dennis Müller

Rather than give a framework of objectives to achieve, we instead introduce a process that can be integrated into the common ways in which mathematical products or services are created, from start to finish. This process helps address the various issues and problems that can arise for the product, the developers, the institution, and for wider society.

To do this, we break down the typical procedure of mathematical development into 10 key stages; our "10 pillars for responsible development" which follow a somewhat chronological ordering of the steps, and associated challenges, that frequently occur in mathematical work. Together these 10 pillars cover issues of the entire lifecycle of a mathematical product or service, including the preparatory work required to responsibly start a project, central questions of good technical mathematics and data science, and issues of communication, deployment and follow-up maintenance specifically related to mathematical systems.

This manifesto, and the pillars within it, are the culmination of 7 years of work done by us as part of the Cambridge University Ethics in Mathematics Project. These are all tried-and-tested ideas, that we have presented and used in both academic and industrial environments. In our work, we have directly seen that mathematics can be an incredible tool for good in society, but also that without careful consideration it can cause immense harm. We hope that following this manifesto will empower its readers to reduce the risk of undesirable and unwanted consequences of their mathematical work.

Lalitha Sundaram, Tom Hobson and Alex Klein contributed to a joint response to the UK's refreshed Biological Security Strategy with colleagues from the Centre for Long-Term Resilience.

Response to the UK Government's refreshed Biological Security Strategy (BSS) on 19 June by Sophie Rose, Cassidy Nelson, Lalitha Sundaram, Tom Hobson, Alexandra Klein and Piers Millett

We are pleased to see many important commitments to strengthening the UK's capabilities for preventing, detecting and responding to biological threats in the Biological Security Strategy (BSS), published on 12 June 2023.

We particularly welcome commitments to formalise the Government's biosecurity leadership, governance and accountability structures, to invest in the UK's real-time biosurveillance and detection capabilities, and to lead internationally in establishing standards of best practice for responsible innovation.

We also commend the Government on allocating £1.5 billion per year to support this work, but urge the Government to continue to sustain a level of investment commensurate with the urgency and importance of implementing the BSS' priority outcomes.

To facilitate the delivery of the Strategy's 15 priority outcomes on such an ambitious timeline, we suggest the Government should:

- Identify reporting milestones and specific, measurable targets for each of the priority outcomes within the Strategy.
- Set out how it will develop thoughtful regulatory standards and practices for ensuring responsible innovation.
- Establish mechanisms for identifying and accessing the diversity of relevant expertise needed to support the Strategy's implementation.
- Ensure a variety of intervention options are being evaluated and appropriately incorporated into future biological event response planning.

Lalitha Sundaram, Tom Hobson and Alex Klein published a report of their workshop following the Ninth Review Conference of the Biological Weapons Convention.

Workshop Report: Ninth Review Conference of the Biological Weapons Convention: Where Next for the UK? on 20 June by Lalitha Sundaram, Tom Hobson and Alex Klein

In April 2023, a group of 19 experts gathered at the University of Cambridge to discuss the outcomes of the Ninth Review Conference of the Biological Weapons Convention, and the implications for biosecurity and non-proliferation in the UK.

The meeting included representatives from:

- academia
- civil society and NGOs
- government and the civil service

Attendees were largely UK based, though the meeting also had

representation from the United States. Together, they brought expertise in:

- biological security policy and implementation
- governance of life sciences research
- non-proliferation and disarmament
- innovation and technology policy

The gathered participants discussed a broad range of issues, but centred on the core issues of:

- recent progress and stagnation at the Ninth Review Conference of the Biological Weapons Convention;
- recent (2018 onwards) efforts within the UK to develop an effective national biosecurity strategy;
- the myriad interactions between national and international fora and mechanisms for biosecurity governance and non-proliferation;
- and those between governments, NGOs, civil society, and practising
- scientists.

Discussions at the meeting ranged from highly pragmatic issues related to the challenges of effective implementation (national and international) and those posed by emerging technologies, through to more foundational conversations about the functional or symbolic nature of different types of formal documentation, policy instruments, diplomatic engagements, and national strategies.

In terms of progress and the (im)possibility of improving or advancing international biosecurity governance, discussions ranged from ambitious and speculative proposals to enhance the meaningful participation of relevant civil society actors and practitioners, through to the realpolitik difficulties of international agreements and diplomacy in specific arenas of negotiation - including the Review Conference itself.

This document provides a summary of the key themes and discussions that took place, aiming to locate them within the context of relevant policy or debate. The report also summarises some key ongoing challenges for those of us in the field.

Shahar Avin published a briefing paper led by colleagues from the Centre for Emerging Technology and Security (CETaS), a research centre based at The Alan Turing Institute, and the Centre for Long-Term Resilience.

Strengthening Resilience to AI Risk: A guide for UK policymakers on 2 August by Ardi Janjeva, Nikhil Mulani, Rosamund Powell, Jess Whittlestone and Shahar Avin

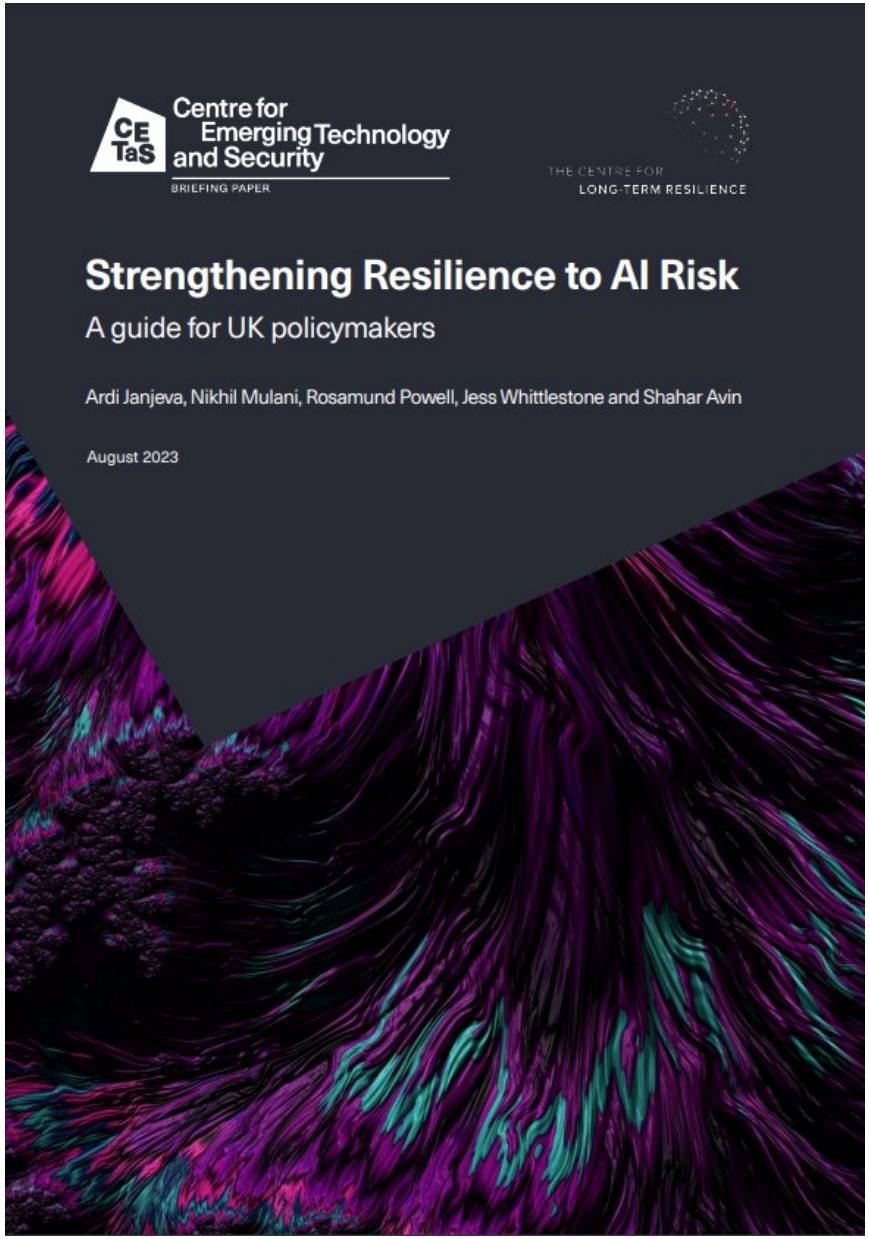
This Briefing Paper from CETaS and CLTR aims to provide a clear framework to inform the UK Government's approach to understanding and responding to the risks posed by Artificial Intelligence (AI). The Government has shown increasing ambition to take a globally leading role in mitigating AI risks, but currently the UK is inadequately resilient to the risks posed by AI. Now is the time to act decisively on the policy interventions required to address those risks.

Any further delay will risk one of two undesirable outcomes: either a scenario where AI risks transition into widespread harms, directly impacting individuals and groups in society; or the converse scenario where widespread fear of AI risk results in a lack of adoption, meaning the UK does not benefit from the many societal benefits presented by these technologies. This paper addresses this challenge by presenting an evidence-based, structured framework for identifying AI risks and associated policy responses.

For the UK to foster a trustworthy AI ecosystem, policymakers must demonstrate both an understanding of and capacity to intervene across the AI lifecycle. This entails addressing risk pathways at their source in the design and training stages, mitigating deployment risks through implementation of clear safeguards, and redressing harmful impacts over the longer-term diffusion of AI systems across society.

The UK is not alone in wanting to mitigate risks from AI while harnessing its wide-ranging societal benefits, in sectors from health and transport to manufacturing and national security. There will be areas of intense geopolitical competition – particularly in research and development capability. But there will also be areas where global cooperation is imperative: the UK cannot safeguard its population from AI risks in isolation, because the harms caused by AI systems do not respect borders. Notwithstanding the critical role of private and third sector stakeholders in shaping the future AI policy landscape, governments must be at the forefront of a global approach which is inclusive, transparent, adaptable, and interdisciplinary in nature.

Future policy must recognise the mutually reinforcing relationship between domestic and global policy interventions: by being proactive in implementing domestic AI policy measures and evaluating their success, the UK will be in a better position to advocate for the adoption of those policies on the global stage, which in turn will generate further support and investment for the UK's domestic AI ecosystem.



Contact

Amanda Lightstone
Head of Development – Arts and Humanities
University of Cambridge 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

UNIVERSITY OF CAMBRIDGE