Wireless Access

Wireless access is possible via UniOFCam or Eduroam (if registered).

If you are not a member of the University of Cambridge and/or do not have Raven or Eduroam access, please collect your username and password at the registration desk.

Acknowledgements

This conference was made possible through the support of a grant from Templeton World Charity Foundation and the Future of Life Institute. The opinions expressed at the conference are those of the presenters and do not necessarily reflect the views of Templeton World Charity Foundation or the Future of Life Institute.

Cover image
A 'Blue Marble' image of the Earth taken from the VIIRS instrument aboard NASA's most recently launched Earth-observing satellite - Suomi NPP.

Image credit: NASA/NOAA/GSFC/Suomi NPP/VIIRS/Norman Kuring
Contents

About CSER ....................................................... 4
Cambridge Conference on Catastrophic Risk .............. 5
Programme .......................................................... 6
Chairs and Keynote speakers ................................. 11
Abstracts and biographies .................................... 12
About CSER

An existential risk is one that threatens the existence of our entire species. The Cambridge Centre for the Study of Existential Risk (CSER) — a joint initiative between a philosopher, a scientist, and a software entrepreneur — was founded on the conviction that these risks require a great deal more scientific investigation than they presently receive.

CSER is a multidisciplinary research centre dedicated to the study and mitigation of risks that could lead to human extinction. Our goal is to steer a small fraction of Cambridge's great intellectual resources, and of the reputation built on its past and present scientific pre-eminence, to the task of ensuring that our own species has a long-term future.

CSER is now hosted within Cambridge’s Centre for Research in the Arts, Social Sciences and Humanities (CRASSH), under the management of Dr. Seán Ó hÉigeartaigh. We received seed funding from founder Jaan Tallinn. Our major research projects, workshops and lectures are supported by the Templeton World Charity Foundation (Managing Extreme Risks), the Grantham Foundation for the Protection of the Environment, the Hauser-Raspe Foundation, and the Blavatnik Foundation (Blavatnik Public Lectures). The Centre’s research is also supported by the Libra Foundation, the Musk Foundation, the Milner Foundation and other generous philanthropic support.
Cambridge Conference on Catastrophic Risk

12 – 14 December 2016
The Gillespie Centre, Clare College, University of Cambridge

The conference is organised by the Centre for the Study of Existential Risk (CSER) and co-sponsored by the Future of Life Institute and the Templeton World Charity Foundation as part of the ‘Managing Extreme Technological Risk’ research programme.

It aims to bring distinct but yet complimentary communities together, to ask how we can best work together and where our efforts should be directed, over the rest of the decade and beyond, in mitigating the most severe potential hazards from human activities. Dedicated sessions, led by invited speakers, will focus on artificial intelligence, depreciation of earth systems and bioengineering. The opening session will give broader consideration to how to better combine our communities’ contributions to efforts for the management of catastrophic risk.

Conference outline

The past five years have seen rapid growth in a thriving set of communities – scientists and technologists in different fields, security and risk researchers, governance specialists, to name a few – united by a common interest in anticipating and mitigating the most severe potential hazards from human activities, particularly from emerging technologies. We still have much to learn from one another, and from the lessons provided by on-going efforts in each of these domains.

Each of the three days of the conference will focus on one of the areas below.

1. Machine Intelligence: Creating a Community for Beneficial AI
2. Depreciation of Earth Systems: Biodiversity, Climate and Environmental Risks
3. Bioengineering: Lessons from Recent Cases for Building Engagement between Communities

Within each focus area we will explore three main themes:

a) Current best understanding of risks and mitigation strategies
b) Lessons from the history of engagement with these risks, in academia, industry and policy
c) Future directions for the communities engaging with these risks.
Programme

Monday 12 December 2016

9:00 – 9:30  Registration / Tea and coffee

OPENING SESSION
(Riley Auditorium)  Chair and Introductory Talk: Huw Price

Keynote Speakers:
Claire Craig
Rowan Douglas

9:30 – 11:00  Opening Session: Part 1

11:00 – 11:30  Break

11:30 – 13:00  Opening Session: Part 2

13:00 – 14:00  Lunch (Garden Room)

SESSION 1  ARTIFICIAL INTELLIGENCE
(Riley Auditorium)  Chair: Marta Halina

Keynote Speakers:
Viktoriya Krakovna
Seán Ó hEigeartaigh
Toby Walsh

14:00 – 15:30  Session 1 – Artificial Intelligence: Part 1

15:30 – 16:00  Break

16:00 – 17:30  Session 1 – Artificial Intelligence: Part 2

18:00  Drinks Reception (Garden Room)
Tuesday 13 December 2016

9:00 – 9:30  Tea and coffee

SESSION 2  DEPRECIATING EARTH SYSTEMS
(Riley Auditorium)

Chair: Partha Dasgupta

Keynote speakers:
Peter Kareiva
Tim Newbold
Tim Palmer

9:30 – 11:00  Session 2 – Depreciating Earth Systems: Part 1

11:00 – 11:30  Break

11:30 – 13:00  Session 2 – Depreciating Earth Systems: Part 2

13:00 – 14:00  Lunch (Garden Room)

14:00 – 15:30  Contributed papers:

Session A – Guiding Technology Development
Elton Room
Chair: (tbc)

Preserving the norm against chemical weapons: an international non-governmental initiative – Malcolm Dando

Ethics management in the human brain project – Bernd Carsten Stahl

The principle of differential technology development – Anders Sandberg

Session B – AI Law and Governance
Bowring Room
Chair: Catherine Rhodes, CSER

The UNW Plan: a realistic political path to beneficial superintelligence – Allan Dafoe

Politics in the age of weapons containment: insights from nuclear weapons policy for the governance of AI development – Matthijs Maas

Complexity Theory as a paradigm for artificial intelligence regulation – Christopher Markou
(cont.)
15:30 – 16:00  Break

16:00 – 17:30  **Session C – Non-technological Risk**  
Elton Room  
Chair: (tbc)

  *Cascading catastrophic risk: future-based approaches to the intersection of disaster and disease* - Charlotte Hammer

  *Extreme solar flares as a catastrophic risk* – Hiroaki Isobe

  *Building resilient local communities: A Franklin County, Ohio case study* – Brook Kohn

**Session D - Catastrophic Risk Methodologies**  
Bowring Room  
Chair: Shahar Avin, CSER

  *Classifying risks of human extinction* - Owen Cotton-Barratt

  *Building a framework for accurate forecasting* - Dan Moinard

  *Lessons from 10 years of public meetups addressing existential risk* – David Wood

18:00  Conference dinner (Great Hall, Old Court)
Wednesday 14 December 2016

9:30 – 9:30  Tea and coffee

SESSION 3  BIO RISK
Chair: Piers Millett

Keynote Speakers:
Jo Husbands
Zabta Shinwari
Sam Weiss Evans

9:30 – 11:00  Session 3 – Bio-risk: Part 1
11:00 – 11:30  Break
11:30 – 13:00  Session 3 – Bio-risk: Part 2
13:00 – 14:00  Lunch (Garden Room)

Contributed papers

14:00 – 15:30  Session E - Policy, Law and Governance
Elton Room
Chair: Julius Weitzdoerfer, CSER

Existential risks as policy problems with legal solutions?
– Kristian Cedervall Lauta

Climate change, catastrophic risks and population policy
– Karin Kuhlemann

The increasing application of criminal law in disasters and tragedies
– Denis Binder

Rights and representation of future generations: challenges for existential risk governance – Natalie Jones

(cont.)
Session F – Managing risks and Benefits
Bowring Room
Chair: Seán Ó hÉigeartaigh, CSER

Solar radiation management, existential risk and moral hazard
– John Halstead

Pricing externalities for dual-use research of concern - Sebastian Farquhar

Gene drive technology: challenges to global risk management and governance – Achim Rosemann

15:30 – 16:00 Break
16:00 – 17:00 Closing panel

Programme details are correct at the time of going to print. Details are subject to change.
Chairs and Keynote Speakers

Claire Craig  
Royal Society

Rowan Douglas  
Willis Group

Jo Husbands  
US National Academies of Sciences, Engineering and Medicine

Peter Kareiva  
UCLA

Viktoriya Krakovna  
DeepMind

Tim Newbold  
University College London

Seán Ó hÉigeartaigh  
University of Cambridge

Tim Palmer  
University of Oxford

Huw Price  
University of Cambridge

Zabta Shinwari  
Quaid-i-Azam University

Toby Walsh  
University of New South Wales

Sam Weiss Evans  
Harvard Kennedy School
Abstracts and biographies

Keynote speakers

Claire Craig

*Risk management in a policy environment: what are the particular challenges associated with existential risks?*

Governments are constantly making decisions that affect risks to society. Contrary to some popular notions, these include the moments when Ministers know their decisions will have extremely significant and long-term effects and when they do their best to take account of them. Governments are also the decision-makers by default at moments of national crisis, when major decisions need to be taken in minutes or hours.

From the perspective of a policy-maker, therefore, existential risks are likely to be viewed through the lens of understanding about risk in policy-making more generally and there is a growing body of experience and insight which informs this. Basic risk frameworks include the need to consider several pathways of action (including inaction), to consider benefits and disbenefits, to consider distributional and transitional effects. They also combine the best insights from the full range of academic disciplines with arrangements to ensure democratic robustness.

In some senses, this rationalist approach pushes the challenges further back and further out. The issue becomes not so much how well specific decisions concerning clearly defined options in a professional environment are taken. The issue becomes questions about what determines the environment that shapes what those options appear to be at any time, and what enables or limits the scope of action of the decision-maker. There is a growing body of work, informed by study of major policy changes over the 20th century, which says that ultimately the case is “not just for a new policy, created by Government, but for a new society created by us all”.

For existential risks where most of the potential pathways are not defined, let alone fully described, there are a further set of challenges to ensuring responsible public and policy debate. However here, too, there is a small but increasing body of knowledge about how individuals and groups can explicitly engage with highly uncertain futures: primarily drawing on the experiences of futures work in the public and private sectors, the evidence from structured public dialogue and new forms of academic insight.

We know, for example, that – not at all unreasonably – in the absence of plausible and evidenced bases for forming views about potential new technologies, people will tend to default to general views about such things as the perceived value of technology as a whole. That they will be concerned with the motives and characteristics of the actors or speakers. That the public debate may be at times path dependent, so the question of which aspects are socially amplified, when and how, may have broader implications for the long-term outcomes.

Dr Claire Craig CBE is Director of Science Policy at the Royal Society. Previously Claire led the Government Office for Science on behalf of the UK Government’s Chief Scientific Advisor (GCSA). She originally joined the Civil Service to run Foresight, a programme of science-based strategic

---

1 Matthew Taylor, Royal Society of Arts, Annual Lecture 2016

12
futures projects covering topics from flood risk to cognitive enhancement. Claire has worked extensively on strategy and science in decision-making. Her career includes periods at McKinsey & Co and the Prime Minister’s Delivery Unit. She helped launch a hands-on science centre in her home town (Bristol), and has held Board roles at a variety of research and teaching institutions including the Council of King’s College London and the Governing Body of Newnham College Cambridge. She trained originally as a geophysicist.

Rowan Douglas
Rowan Douglas leads the Capital, Science and Policy Practice at Willis Group. The Practice confronts large-scale challenges of risk, resilience and sustainable growth at global and local scales through public, private and mutual mechanisms. Willis Group is an insurance and reinsurance broker with approximately 20,000 personnel operating in around 100 countries.

Rowan has two public appointments in the UK serving on the Prime Minister's Council for Science and Technology and also the Natural Environment Research Council, which oversees approximately $500m of annual environmental science expenditure.

More widely, he Chairs the United Nations International Strategy for Disaster Reduction Private and Financial Sector Working Group, preparing the second UN Hyogo Framework for Action Agreement in 2015 and the World Meteorological Organisation Expert Advisory Group on Financial Risk Transfer, preparing for the UN Agreement on Climate Services in 2015. He is also a member of the Political Champions for Disaster Reduction Committee, chaired by the UNDP Secretary General and the UK Secretary of State of International Development.

He is a member of the Global Earthquake Model Foundation Governing Board, Pavia Italy; the Advisory Board of the Earth System Laboratory, NCAR, Boulder CO and the Royal Society’s Working Group on Human Resilience to Climate Change.

Rowan read Geography at Durham University (BA Hons); Geographical Sciences at the University of Bristol (MPhil) and is a Fellow of the Royal Geographical Society.

Jo Husbands
Competing catastrophes: lessons from efforts to mitigate the risks of “dual use” research
The presentation will focus on lessons from efforts over the last 15 years to engage scientists in the development of strategies to mitigate the potential risks of continuing rapid advances in the life sciences. The primary focus has been on potential oversight mechanisms for “dual use” research. In the case of the life sciences, dual use refers to research that, although intended for beneficial purposes, has the potential to be misused for bioterrorism or for bioweapons. The premise behind such efforts is that effective policy and practice will require active commitment by scientists, who are part of many relevant sectors (academia, industry, public health and government) and that the scientific community has important roles to play. These roles include conducting education and outreach programs as well as helping to inform policymakers about developments in science and technology and their implications for security.
Discussion and debate over how “biosecurity” is best framed continues to be a key part of efforts to develop an engagement strategy. The presentation argues that the framing must:

- engage many stakeholders and fit within mandates of relevant governments and international organisations;
- facilitate reaching widest range of scientists (e.g. in academia, industry, public health);
- be compatible with more security-focused activities for specialised, more directly affected audiences; and
- complement the existing legal and regulatory structure and provide a basis for discussing additional measures or changes in practices.

The presentation argues that the most effective framing places biosecurity within the broader context of the responsible conduct of science.

To illustrate this argument, the presentation will develop two cases. One is the activities of the InterAcademy Partnership (IAP) and its Biosecurity Working Group, in particular the relationship the Working Group has built with the Biological Weapons Convention. It will also illustrate how the current structure of the BWC has provided an important mechanism that supports the efforts of IAP and other scientific organisations to engage scientists in biosecurity issues.

The second case is the controversy that erupted in late 2011 over so-called "gain-of-function" (GOF) research and continues to incite periodic debate. The emphasis is on the role played by the U.S. National Academies of Sciences, Engineering, and Medicine in convening public symposia to provide a mechanism for public engagement in a White House led process to develop policy for the effective oversight of research. The case also includes the role of the European Academies Science Advisory Council in a study it conducted at the request of the European Commission to provide recommendations for EC policy. One of the features of the GOF controversy was the division within the scientific community over the potential risks and benefits of the research.

Taken together, these cases provide the basis for discussing “our current best understanding of risks and mitigation strategies; lessons from the history of engagement with such risks; and future directions for the community engaging with their management.”

**Dr Jo Husbands** is a Scholar with the Board on Life Sciences of the National Academies of Sciences, Engineering, and Medicine, where her work focuses on biosecurity and dual use issues. She concentrates, first, on education and outreach in the broader context of responsible science and, secondly, on the implications of continuing advances in the life sciences for efforts to mitigate the risks of misuse. She also represents the Academies on the Biosecurity Working Group of the InterAcademy Partnership (IAP). From 1991 – 2005 she was Director of the Academies’ Committee on International Security and Arms Control and its Working Group on Biological Weapons Control.

Before joining the National Academies, she worked for several Washington DC-based nongovernmental organisations engaged with international security. From 2001 – 2012 Dr Husbands was an adjunct professor in the Security Studies Program at Georgetown University. She is a Fellow of the International Union of Pure and Applied Chemistry and an inaugural member of the Advisory Board on Education and Outreach of the Organization for the Prohibition of Chemical Weapons. She holds a PhD in Political Science from the University of Minnesota and a
Masters in International Public Policy (International Economics) from the Johns Hopkins University School of Advanced International Studies.

Peter Kareiva

*Why we are unlikely to foresee an environmental apocalypse, and what to do about it*

Famine and pestilence are common Malthusian narratives that link human avarice and fertility with planetary catastrophe. These Biblical judgement days have been with us for centuries, with little supporting evidence. Equally unreal are the counter-narratives of blind techno-optimism, whereby whatever problem arises, is surely solved by human ingenuity – so why worry? What then are the environmental catastrophes that could threaten human existence?

1. The first is no surprise: global climate disruption. Environmental scientists, national security experts, and global businesses all recognise climate change as a potentially catastrophic threat. Society, businesses, and governments are taking action but the risk is that positive feedbacks will accelerate warming and proceed so rapidly that efforts to reduce emissions happen too late.

2. After climate, loss of soil looms as a potential global disaster. Currently soil is being lost at ten times the rate it is being generated, which prompted the UN to declare 2015 to be labelled the “International Year of Soils”.

3. Emerging diseases represent a third disaster waiting to happen. HIV jumped from monkeys to humans, and has killed 35 million people since the first diagnosed case in 1959. HIV has a low transmission rate — the worry is that something with a much higher transmission rate might jump to the human species.

All of these ominous risks are exacerbated by an absence of modularity, and the highly nonlinear and strong interactions that reverberate throughout human-natural systems. Hence, a local famine can produce a massive exodus of people who then destabilise either the ecosystem or political system into which they move. Alternatively, in the absence of adequate migration, the local disaster becomes so severe that instead of migrating people exporting stress, dust or pollutants are transported globally.

Especially disturbing is the possibility that environmental science is ill equipped to foresee existential threats. Three scientific shortcomings expose the world to environmental catastrophe. First, there is a tendency to overstate certain “popular” threats and thereby inadvertently divert attention away from the real risks. Second, in the face of extreme change and environmental shocks, we show little ability to manage for change, or to take radical action in the face of severe degradation. Instead there remains a tendency to preserve ecosystems in historical states in direct opposition to the changing world around us, and to be locked into rigid regulatory approaches as opposed to experimental management. And the last and perhaps most ominous environmental threat is our ignorance of ecosystems, and our failure to invest in collecting or using data that might provide real-time signals of potential disasters. Instead, the environmental and conservation community focuses on obscure indices, and expert opinions that are infrequently updated. It is analogous to a medical doctor taking no physiological measurements, but assessing one’s health by looking at family photos and talking with several family friends to construct mysterious indices of health.
Peter Kareiva is Director of the Institute of the Environment and Sustainability at UCLA. Prior to this, he was Chief Scientist for The Nature Conservancy, the world’s largest international science-based environmental non-profit. He chairs the Board of the Science for Nature and People Partnership (SNAPP), a collaboration among The Nature Conservancy, the Wildlife Conservation Society and the National Center for Ecological Analysis and Synthesis that is designed to rapidly respond to critical questions involving nature and human well-being. Kareiva studied political science and zoology at Duke University for his bachelor’s degree and ecology and applied mathematics at Cornell University for his PhD.

Peter is author of more than 150 scientific publications and author or editor of eight books, including a textbook on conservation science. His current research concerns the connection between human activities and changes in ecosystem services, as part of the Natural Capital Project. Kareiva is also studying the linkage between the sustainability initiatives of global corporations and their impacts on ecosystems, as well as the value of nature for people in urban areas. He is a member of the National Academy of Science and a Fellow of the American Academy of Arts and Science.

Viktoriya Krakovna
AI risk: past, present, and future
In the past few years, we have seen a dramatic increase in discussion and engagement on AI safety issues in the AI research community. I will draw some lessons from the efforts that helped to bring AI safety into the mainstream, discuss the current state of the field, and consider outreach strategies and research priorities going forward.

Viktoriya Krakovna is a research scientist in AI safety at DeepMind and co-founder of the Future of Life Institute. Her PhD thesis in statistics and machine learning at Harvard University focused on building interpretable models. Viktoriya holds numerous distinctions for her accomplishments in mathematics competitions, including a silver medal at the International Mathematical Olympiad.

Tim Newbold
Global effects of land use on the diversity of local ecological communities
Human use of the land presents the greatest threat to biodiversity currently. Until recently, knowledge about the effects of land use on biodiversity has been restricted to certain parts of the world and to well-studied groups of species. I will present results from the PREDICTS Project, which has gathered over 3 million records for over 40,000 species, in different land uses, at 27,000 sites around the world. I will talk about the use of the PREDICTS data to generate models of how biodiversity responds to land use, and the application of these models to scenarios of land use to predict the current status and potential futures for biodiversity. I will also discuss our recent effort to relate our estimates of biodiversity loss to the proposed planetary boundary for biodiversity, beyond which the ability of biodiversity to support critical ecosystem functions might become uncertain.

Tim Newbold is a Royal Society University Research Fellow, based in the Centre for Biodiversity and Environment Research at University College London. His research interests are in
understanding and predicting the effects of human activities (especially through land use and climate change) on the structure and diversity of ecological communities. Before moving to University College London, Tim was based at the United Nations Environment Programme World Conservation Monitoring Centre, where he also helped develop the first global General Ecosystem Model, the Madingley Model.

Seán Ó hÉigeartaigh
Seán Ó hÉigeartaigh is the Executive Director of Cambridge's Centre for the Study of Existential Risk (CSER), an academic research centre focused on global catastrophic and existential risks associated with emerging technologies and human activity. In the last five years, he has played a central role in research projects on long-term trajectories and impacts associated with artificial intelligence (AI) and other emerging technologies, project research programmes at the Future of Humanity Institute (Oxford) from 2011 – 2015, co-developing the Strategic AI Research Centre (Cambridge-Oxford collaboration) in 2015, and the Leverhulme Centre for the Future of Intelligence (Cambridge-Oxford-Imperial-Berkeley collaboration) in 2015/16. Seán's research spans technology policy and strategy, catastrophic risk, and horizon-scanning and foresight.

Tim Palmer
Climate change: catastrophe, hoax or just lukewarm?
In a recent talk at the Royal Society, climate sceptic Matt Ridley proposed a novel trichotomous characterisation of the climate-change problem, as outlined in the title above. He described himself as a "lukewarmer". However, by virtue of the complex nonlinear nature of climate, such a categorical delineation of the problem is inappropriate. Rather, to be consistent with the nature of key uncertainties in estimating how increasing levels of atmospheric carbon dioxide interact with the natural climate system, notably with the hydrological cycle, estimates of future climate should instead be framed in terms of probability and risk: What, for example, is the risk of catastrophic climate change (the sort that would pose an existential threat to significant sections of humanity) in the coming century or so? It is simply inconsistent with our present state of knowledge to conclude that this risk is negligible. Indeed, according to both observations and modelling, this risk lies in the 10s of per cent, and is therefore substantial by any relevant measure.

The biggest uncertainty preventing a more categorical estimate of future climate lies in the problem of cloud feedback: How will low- and high-level cloud respond to increasing levels of carbon dioxide? Current computers are not powerful enough to allow clouds to be resolved and they must therefore be represented by quasi-empirical parametrisation formulae. Here I call on the climate sceptic community (including, perhaps, the incoming US Administration) to acknowledge this risk-based approach to the assessment of future climate and to broadly accept IPCC estimates of risk (whilst recognising that these estimates can and must be improved). I call on economists and climate scientists to work closer together to provide the most reliable estimates possible of expected climate impacts. To aid such work, I call on European countries to come together to fund a dedicated (exascale) flagship climate-computing centre – a sort of Climate CERN – which would allow the resolution of global climate models to be driven down to scales where key cloud systems can be represented using the laws of physics. Such a centre is essential if we are to understand and quantify better the nature of climate extremes.
**Tim Palmer** FRS CBE is a Royal Society Research Professor at the University of Oxford. He has contributed to all IPCC reports and chaired the science steering group of the World Climate Research Programme’s committee, which provides the key modelling input to IPCC. He is best known for his work developing ensemble-based probabilistic forecast methods for weather and climate prediction – a technique which has transformed operational weather and climate prediction around the world over the last decade or so. He has won a number of prizes, including the top prizes of the European and American meteorological societies and the 2014 Dirac Gold Medal of the Institute of Physics. He is a Fellow of a number of learned societies around the world. Tim has been especially active in European collaborative research on weather and climate problems where he helped develop links to downstream applications such as malaria and crop-yield prediction. His PhD was in general relativity theory at the University of Oxford where he developed the first quasi-local expressions for gravitational energy-momentum. He remains active in the field of fundamental physics.

**Huw Price**
Huw Price is Bertrand Russell Professor of Philosophy and a Fellow of Trinity College, University of Cambridge. He is also Academic Director of the Centre for the Study of Existential Risk, which he founded in 2012 with Martin Rees and Jaan Tallinn. He is also Academic Director of the new Leverhulme Centre for the Future of Intelligence. Before moving to Cambridge he was ARC Federation Fellow and Challis Professor of Philosophy at the University of Sydney, where from 2002 – 2012 he was Director of the Centre for Time.

**Zabta K. Shinwari**

*Engagement of young researchers in responsible conduct of science: successes and failures*

We scientists, in general, practise science for the benefit of humanity. However, there is a possibility of the misuse or dual use of science, not only by scientists themselves but by non-state actors as well. Some inventions may harm the environment, or society in general, hence researchers must be aware of the rights of individuals and of populations, and have a sense of responsibility to the wider world. In a way, a Responsible Conduct of Science requires double consideration: the intention of scientists is, on one hand, the advancement of science and its positive impact on society and, on the other the reporting of scientific inventions to peers and society in a way that no harm comes out of their scientific exploration.

In this era where new technologies are emerging, pathogens are also emerging and re-emerging, and laws and social norms can deviate from what is ethical. Therefore, scientists have to constantly examine their standards to ensure that they are reasonable and well founded. Scientists have to use a more teleological ethics, which argues that the morality of an action depends on its outcome.

There should also be ethical considerations around falsification, fabrication and plagiarism. Scientists have a responsibility to not only carefully avoid and detect unforeseen or unintended events, but to make other stakeholders, including politicians, aware of how to avoid or minimise the threats stemming from dual use of science. We also need awareness in some fields, such as biotechnology, which lack accurate, definite predictions of outcomes.
Society expects "Maximizing wellbeing and minimizing pain of humanity" as a result of science. The aim of biotechnology should be to reduce human suffering, disease, hunger and poverty. To adopt this technology on a large scale, we need to overcome the hurdles, which are largely political in nature. Bioethicists, biotechnologists, political activists, NGOs and other actors have to find common ground on how to address the challenges raised by biotechnology with regard to the respect for human dignity and economic inequalities between and within societies.

Another aspect of the responsible approach to biotechnology determines that we must use safe and environment-friendly techniques when creating new organisms. Such slogans as "From Reading to Writing the Genetic Code" led to the designing of synthetic organisms, but they have also raised ethical questions about scientists "playing the role of God" and more generally about the relationship between science and religion.

Ignoring science is the most unethical attitude. However, science requires critical examination of the matter and open dialogue. It does not tolerate doctrinaire visions. This is why consideration of many ethical issues, such as the elimination of inequalities between human beings and ethical challenges raised by modern biotechnologies in many developing countries, continues to remain a challenge.

The so-called "dual-use dilemma" arises in the context of research in the biological and other sciences because the same piece of scientific research sometimes has the potential to be used for harm as well as for good. Research in biotechnology and other life sciences has massive potential to be used for nefarious purposes The recent progress in biotechnology has offered tremendous opportunities globally, but as one of the most rapidly-growing areas of science in the early 21st century, it brings security risks that must be recognised and addressed effectively. So it is of critical importance that the scientist has substantial knowledge and understanding to cope with these biosecurity risks. Our work has shown that scientists’ knowledge and understanding of these issues is largely non-existent without the implementation of biosecurity education.

**Dr Zabta Khan Shinwari** (PhD, T.I. (Fellow-PAS)) is currently Professor and Chairman of Biotechnology at Quaid-i-Azam University. He was Vice Chancellor of Kohat University of Science & Technology, established the University of Science & Technology in Bannu and served as CEO of Qarshi Research International and Vice Chancellor of Qarshi University-Lahore. He is the founder of KUST Institute of Medical Sciences (KIMS).

Dr. Shinwari’s interests include Molecular Systematics, Bioethics and Biotechnology. He is especially interested in teaching and developing Modules on Dual Use Education. He has been a partner in promoting Biosecurity education in Pakistan, as well as being a key figure in terms of Biosecurity and international engagement on Biosecurity issues. He has published extensively both in Pakistan and internationally.

**Toby Walsh**

**The Limits of AI**

As we see breakthroughs every day in the field of Artificial Intelligence (AI), the question arises as to what are the limits to this scientific quest? Will we re-create human or super-human level intelligence in machines? Will they be conscious? Will machines take over as the dominant species on the planet? Where will this all end?
Criminal law serves two purposes: to punish and to deter. Punishment is the imposition of justice and retribution. Deterrence sends a message that society does not tolerate this conduct. Those who engage in it risk punishment.
Civil liability is intended to compensate the victims, but may not always deter. Damages may be small, the costs of litigation great, insurance or workers compensation may cover the costs, the jurisdiction may limit liability, or the defendant may be judgment proof, such as through bankruptcy. Thus, criminal liability may be the only effective means for society to send the message that this conduct is unacceptable.

Institutions, such as corporations and governments, can only act through individuals. Someone made the decision to act or not act that resulted in injury. Vicarious liability usually shifts the costs to the employer such that the wrongful actor may escape liability, although the actor technically remains personally liable.

Criminal prosecutions may be brought against both the employer and the employee. The risks of arrest, booking, jail, bail, prosecution, conviction, prison time, probation, fines and humiliation affect the defendant much more than the risk of civil liability. Corporations cannot be imprisoned, but they can be fined, placed on probation, and have special monitors appointed to oversee them. Corporations and individuals also run the risk of ignominy from criminal prosecutions.

This study tests four theses:
1. criminal prosecutions are increasing in disasters and tragedies,
2. the prosecutions occur globally, but are especially prevalent in Asia and the Pacific Islands,
3. government employees and officials are often prosecuted for corruption and dereliction of duty, and
4. the Digital Age with Smart Phones, Tablets and Social Media is fuelling the rise in prosecutions.

The only way to test these theses was through empirical research to compile a comprehensive list of incidents. Lists existed for specific risks, such as aviation disasters, but no comprehensive list existed until now.

The study is on going as incidents arise or previous cases are uncovered. It is limited by the information available online though the Internet in English. The final resolution of the cases is not always available.

Professor Denis Binder received his doctorate (S.J.D.) from the University of Michigan in 1973. He has been teaching law for 45 years, mostly Environmental Law, Torts, and Toxic Torts. He became involved in infrastructure issues, especially the legal aspects of dam safety, in 1976, followed by emergency action planning and business continuity after the 9/11 terrorist attacks in the United States, and now campus security and cyber security. He received the National Award of Merit in 1996 from the Association of State Dam safety Officials.

Professor Binder’s perspective is legal, which entails an analysis of the causes of disasters and tragedies. He has written and spoken extensively on issues of safety and emergency planning, both in the United States and globally.
Kristian Cedervall Lauta and Hin-Yan Liu

Existential risks as policy problems with legal Solutions?

Despite the significant anthropogenic component of many, if not most, existential risks policy and governance tools have remained under-theorised, under-developed and under-deployed. In this paper, we propose that governance perspectives remain marginalised as a result of the technological dominance in the processes of identifying and addressing existential risks. We suggest that perilous blindsides are created insofar as this technological bias crowds out converging discourses, both because potential existential outcomes elude identification and response, and because of the overreliance placed on a narrow set of technological tools to meet the challenge.

Our contribution to redressing this imbalance views the landscape from the law and policy vantage point. This perspective leverages three main benefits, which help to reach towards an equilibrium with technological approaches. First, we reverse our focus: where the existential risks paradigm approaches the problem of human extinction by identifying the range of possible causes, we propose instead to emphasise the impact in order to broaden out the range of potential paths to that end. In doing so, we suggest that the field of existential threats is considerably broader than the current range of existential risks, since human extinction can be precipitated from events and trajectories falling short of the extreme category. Second, we introduce the dimensions of vulnerability and exposure from disaster research to counterbalance the prevailing focus on hazards. This allows us to formulate a typology to identify areas where the characteristics are particularly amenable to law and policy responses. Finally, we depart from what might be under the broader umbrella of negative responses to existential threats that are characterised by a defensive orientation, and move towards a positive sphere of enabling and empower responses epitomised by the production of global public goods. Rather than attempt to directly chisel at specific existential risks, this more general approach suggests more general responses that tangentially diffuse such threats.

We hope that this paper offers a broader, complementary, approach to the technologically grounded focus on existential risks that illustrates the contribution of other fields to protecting the future of humanity. As such, our aim is to open the field to those working in other disciplines to consider how their perspectives might usefully be brought to bear on these grand challenges.

Given the scale and scope of existential risks, fostering the engagement of broader disciplinary perspectives is prudent and necessary. To avoid depreciating the potential contribution of other fields to addressing existential risks, however, processes that convert the complex challenges into terms and concepts that can be addressed within disciplinary confines need to be established. The overarching aim is thus to stimulate disciplinary engagement with existential risks on their own terms: for disciplines to develop typologies capable of identifying the categories of risks that their tools might usefully address independently.
Kristian Cedervall Lauta is Associate Professor at the Center of International Law, Conflict and Crisis, University of Copenhagen (UCPH). His research regards law and disasters, and his latest book, *Disaster Law*, investigates the emergence of an international field on law and disasters. Kristian is part of a number of interdisciplinary and international projects, and is presently co-heading the interdisciplinary research project Changing Disasters. Furthermore, he is actively involved in the organisation of inter-institutional research center COPE; in UCPHs Sustainability Science Center; in the Nordic Centre of Excellence NORDRESS; and in H2020-project ESPRESSO.

Hin-Yan Liu is Associate Professor at the Centre for International Law, Conflict and Crisis, at the University of Copenhagen. His work explores the structural inadequacies of law and systemic shortcomings policy in relation to new actors and emerging technologies. This is reflected in his latest book, *Law’s Impunity* (Cambridge University Press), and his co-leadership of the Autonomous Weapons Systems project organised out of the European University Institute. He has held academic positions at the University of Westminster, King’s College London and NYU Florence, and has been a Fellow at the European University Institute and the Max Planck Institute for Foreign and International Criminal Law.

Owen Cotton-Barratt

*Classifying risks of human extinction*

Risks of human extinction are a major category of existential risk. In this work, we look at ways of slicing up the space of extinction risks according to how the risk process interacts with people or social systems at different points along its trajectory.

The motivation for this work is two-fold:

1. To work towards a common framing which will allow study of extinction risks to remain more of a coherent field, even while much of the research is specialised on particular risk.
2. To better understand and inform decisions about how to prioritise between work trying to reduce different risks and at different stages of the process.

We slice the space three ways. First, by how the process starts in a damaging way. This extends the traditional natural vs. anthropogenic split by subdividing anthropogenic according to how many people are crucially involved, and their level of knowledge and intentionality about the harms caused. Second, by why it is not stopped when it becomes clear it is causing significant damage. Third, by how it can reach enough people to cause extinction.

We then look at possible work to reduce extinction risk. To avoid extinction, a process must be stopped at one of these three steps. Generically, if we have reasonable methods for reducing risk of passing each stage, diminishing returns on action mean we should pursue all of these in parallel rather than committing to our single best guess. This is amplified when prevention work at a stage is cross-cutting and affects multiple risks. An exception may be if we are extremely confident that a risk will get past a given stage, if it reaches it.

Considering our risk-reducing processes also highlights their importance. This gives a possible framework for analysing compound risks, where one global catastrophe, which by itself does not pose an existential risk, might damage our capacity to respond to subsequent risks. The net
existential risk created by a catastrophe is a function of the degree of damage to future ability to lower risks, the persistence of this damage, and the background rate of risk-creating processes.

Owen Cotton-Barratt is a Research Fellow at the Future of Humanity Institute at the University of Oxford, specialising in decision-making under uncertainty and cause prioritisation.

Allan Dafoe

The UNW Plan: A realistic political path to beneficial superintelligence

Machine superintelligence could plausibly arrive in the coming decades. As it approaches, humanity will confront tremendous global political challenges related to the need to prevent a dangerous AI race, to have the option to control dangerous technology throughout the world (“stabilization”), and to have the option to establish a global regime for making deliberate, safe, transparent, inclusive, beneficial choices about our collective future. Envisioning a realistic robust path to solve these challenges is of great value. This paper attempts to sketch such a path that succeeds by embracing the occurrence of a dangerous AI race between great powers. Along the way it identifies and evaluates several crucial forces, trade-offs, and dynamics.

Allan Dafoe is an Assistant Professor of Political Science at Yale University and a Research Associate at the Future of Humanity Institute. His research examines the causes of great power war and the global politics of artificial intelligence. To help scientists better study these and other topics, he also works on methods for causal inference and for promoting transparency.

Malcolm Dando, Michael Crowley and Lijun Shang, University of Bradford

Preserving the norm against chemical weapons: an international non-governmental initiative

As Professor Matthew Meselson pointed out in his 2000 essay Preventing the Hostile Exploitation of Biotechnology, during this century we will increasingly be able to manipulate life “including the processes of cognition, development, reproduction and inheritance.” Meselson also noted that unlike the technologies of nuclear and conventional weapons, “biotechnology has the potential to place mass destructive capabilities in a multitude of hands.”

The key restraints on such appalling possibilities are the norms prohibiting development, possession and use of chemical and biological weapons embodied in the 1997 Chemical Weapons Convention (CWC) and the 1975 Biological and Toxin Weapons Convention (BTWC). The operation and future development of these Conventions, and the norms that they embody, are subject to five-yearly reviews. The BTWC 8th Review Conference will take place in late 2016 and the 4th Review Conference of the CWC will be in 2018.

The project described in this paper is designed to provide an input from non-governmental sources to the 2018 review of the CWC. The project began in 2015 and involves the production of a book that will be published by the UK Royal Society of Chemistry. The book has been written by a wide range of international scholars, life and chemical scientists and experts in arms control and disarmament, and has three major sections in addition to the introduction and conclusion.
The sections cover:
1. the chemical and biological weapons prevention and disarmament regime;
2. relevant advances in chemistry, biology and relevant science and technology such as nanotechnology; and
3. implications for arms control and disarmament – how can the international governmental and non-governmental communities effectively combat the potential emergence, proliferation and use of new chemical weapons.

The overall aim of the project is to provide a detailed background briefing for a wide audience about what is at stake in the 2018 review of the CWC. In short, this paper provides a case study of how international scholars, life and chemical scientists and experts in arms control and disarmament can work together to enhance governance of extreme technological risks.

Professor Malcolm Dando trained originally as a biologist (BSc and PhD at the University of St Andrews, Scotland). After post-doctoral studies in the United States (University of Michigan and University of Oregon) he held UK Ministry of Defence funded fellowships in Operational Research at the University of Sussex during the 1970s. Since then he has worked on arms control and disarmament, particularly on chemical and biological issues (DSc. University of Bradford). In recent years this work has been focused on awareness raising and the education of life scientists in regard to dual use and biosecurity, for example in the Royal Society Brain Waves module on Neuroscience, conflict and security. Professor Dando’s current research is concerned with the implications of the search for novel incapacitating agents on the future of the chemical and biological disarmament regime, for example in his 2015 book Neuroscience and the Future of Chemical-Biological Weapons (Palgrave/Macmillan, Basingstoke).

Sebastian Farquhar, Owen Cotton-Barratt and Andrew Snyder-Beattie

Pricing externalities for dual-use research of concern

How should funders of research evaluate dual-use research of concern? Some risky work is valuable, but accepting too much risk may be ethically neglectful. In this paper, we outline an approach for pricing externalities into the cost of grants for dual-use research. While benefits of research are often very hard to evaluate quantitatively, risks estimates are sometimes more tractable. By pricing these risks and incorporating them into the cost of the grant, funders could then evaluate grants as usual - comparing the scientific merits of a grant against its full cost.

This offers four main advantages. First, it aligns incentives by assigning costs to the generator of the externality. Second, it keeps research funding decisions in the hands of scientists. Third, it incorporates democratic or public input on the significance of risks. Fourth, it ensures appropriate risk assessment and governance expertise is brought to bear. However, it is still hard to estimate the sizes of externalities, and the mechanisms used could create unintended consequences. In order to be revenue neutral for the sciences as a whole, fields would need to receive offsetting funding increases.

Throughout this paper, we use gain-of-function research of concern as a case study. In the first section, we outline the current context and discuss the difficulties of existing approaches to evaluating risky research. In the second section, we outline a general framework for costing externalities of research. In the third section, we describe an insurance market-based and
regulator-based approach for estimating these externalities. In the fourth section, we compare these two approaches against each other and the status quo.

**Sebastian Farquhar** is a researcher at the Centre for Effective Altruism, focusing on policy interventions responding to catastrophic risk scenarios.

**Owen Cotton-Barratt** is a Research Fellow at the Future of Humanity Institute at the University of Oxford, specialising in decision-making under uncertainty and cause prioritisation.

**Andrew Snyder-Beattie** is the Director of Research at the Future of Humanity Institute at the University of Oxford.

**John Halstead**  
**Solar radiation management, existential risk and moral hazard**  
Solar Radiation Management (SRM) is a controversial form of geoengineering which would cool the planet by reflecting sunlight back to space. The most discussed and researched form of SRM involves the injection of aerosols into the stratosphere. SRM has some unique advantages over more conventional ways of reducing climate risk, such as mitigation, in that it is the only known way to quickly and cheaply reduce global temperatures. However, it also brings major risks. This paper discusses whether SRM should receive further research funding, from the point of view of those concerned with reducing existential risk.

One salient concern about SRM research is that it is a *moral hazard risk*. I argue that the moral hazard concern is best dealt with by David Morrow’s framework, according to which SRM research is a *pernicious* moral hazard if and only if it reduces the overall amount of mitigation in the climate policy portfolio, and results in a policy portfolio that is worse overall than the one which would otherwise have been used.

From the point of view of existential risk reduction, SRM research presents a risk-risk scenario: both SRM research and the failure to do SRM research could arguably increase existential risk. Therefore, principles such as Maxipok and the precautionary principle do not provide any clear guidance. Nonetheless, I argue that in spite of the large uncertainty about the question, further SRM research is justified. The main reason for this is, as others have argued, that there is plausibly a high chance that the rest of the policy portfolio will be inadequate regardless of whether SRM is further researched. I make this argument more precise by providing a probability estimate that SRM will be used even if it is not further researched. Firstly, due to political inertia, it is unlikely (<50% chance) that the world will keep to the 2°C target required by the Paris Agreement without deploying SRM. Secondly, even assuming that the international community will ignore the Paris Agreement, a conservative median estimate of the chance of eventual SRM deployment is 15% (with large uncertainty). This has two implications. Firstly, the expected impact of SRM research on the rest of the policy portfolio is relatively low. Secondly, if the incentives to deploy SRM in the future (even if it is not researched) will be so strong, it would be reckless not to research SRM further. Doing so would leave future generations with an under-researched technology on which the lives of billions would depend.

I conclude by outlining ways in which scientists and funding bodies can reduce moral hazard risk.

26
**John Halstead** is currently a research assistant at Somerville College, Oxford, and a researcher on Dave Denkenberger’s project on managing severe food supply shocks. He has a DPhil in political philosophy from St Anne’s College, Oxford, and previously taught ethics to Masters students at the Blavatnik School of Government. He has published on topics such as aggregation, decision theory, and democratic theory in leading philosophy journals including *Ethics, Utilitas, and Ethical Theory and Moral Practice*. After leaving academia, he worked as a research fellow at the Global Priorities Project. He is a member of the effective altruism community and hopes to use his time to reduce existential risk.

**Charlotte Christiane Hammer**

*Cascading catastrophic risk: future-based approaches to the intersection of disaster and disease*

This research seeks to draw on the unique perspectives of risk and future based approaches in order to understand the association between catastrophic flood risk and infectious diseases and toxic exposure. While the interaction of these two colliding fields has been studied in great detail in the context of low-income countries, a lack of knowledge concerning their relationship in industrialised countries can be observed. Furthermore, the relationship is often characterised by either too little or too much conflicting data and thus decision-making is complicated. The research seeks to explore the usefulness of pre-emptive and possibilistic approaches, as they are used in the field of security risks, for the management of situations in which biosecurity consequences of catastrophic flooding occur. While mechanisms of futurity feature prominently in the discussion about terrorism and asymmetric warfare, they are still underdeveloped for other fields of risk, such as disaster related biorisks. Using the examples of the 2002 August river flooding in Germany and Hurricane Katrina, the research examines the applicability of these approaches for such a scenario as a first step towards integrating risk-based, future oriented approaches in other fields of risk, based on the context of emergent risks and risk mobility.

**Charlotte Hammer** is currently a PhD candidate at the University of East Anglia's (UEA) Norwich Medical School. Prior to coming to UEA, she worked in humanitarian aid and research. She holds a Master of Arts in Risk and Security from Durham University and a Bachelor of Arts in Politics and Society from the University of Bonn in Germany. Her primary research interests include biosecurity, disaster risk reduction and management, communicable disease control, and complex humanitarian emergencies.

**Hiroaki Isobe**

*Extreme solar flares as a catastrophic risk*

The sun is a variable star. The sudden release of magnetic energy stored around the sunspots, called solar flares, emit intense electromagnetic radiation, high-energy particles and huge amount of magnetised plasma into the interplanetary space, and sometimes parts of them surge toward the Earth. They were not considered to be hazardous until recently, as there was no notable consequence except for auroras. However, modern society is becoming increasingly vulnerable to the solar variability as it relies more and more on the space assets and the large-scale power grids, which suffers from the disturbances originated from the sun. The most intense solar flare ever known is the so-called Carrington flare, observed by the English astronomer R.C. Carrington in 1859, which caused a very strong geomagnetic storm and low-latitude auroras over the world.
It is estimated that, if the same event occurs and strikes the Earth now, it will cause devastating damage to the satellites and ground facilities, taking many years and trillions of dollars to restore. Moreover, recent studies have shown that even more catastrophic disaster may occur due to the extremely intense solar flares. A group of researchers and students at Kyoto University analysed the photometric observations of many stars whose properties are similar to those of the Sun, and found numerous “superflares”, whose energy exceeds $10^{28}$ J, 1000 times larger than that of the Carrington event. If such superflares occur in the present sun, the consequence will be catastrophic.

In this paper, we review the recent advancements of the studies of extremely intense superflares in the sun-like stars, and discuss the possible consequences of such superflares in the sun. They include significant depletion of the ozone layer for years, almost complete destruction of the space assets, and hazardous radiation dose to astronauts and even to aircraft passengers. We also present the results of our survey of the records of law-latitude auroras in the historical documents, mainly from East Asia but also from the West. The historical aurora records in law latitude regions provide an independent proxy for the occurrence of extreme space weather events. Moreover, by examining the history of people’s responses to natural disasters and the anomalous events such as aurora, we can infer how scientific thinking and preparedness for natural disasters have been fostered among people. Their implication to the present society will be discussed.

Hiroaki Isobe is an associate professor at GSAIS, Kyoto University. He has been working on solar physics and space weather at Kyoto University, University of Tokyo and University of Cambridge. In addition to his main research on the sun, he has led several interdisciplinary research projects such as ethical and anthropological studies of space exploration. He is also engaged in Japanese space policy, public engagement of science, and dialogue between science and Buddhism.

Natalie Jones, Mark O’Brien and Thomas Ryan

*Rights and representation of future generations: challenges for existential risk governance*

Unprecedented technological risks such as artificial intelligence, synthetic biology and distributed manufacturing raise distinctive governance challenges when considering the rights and representation of future generations. The governance of such risks is an intergenerational public good, and in general interests of future generations are under-represented in the status quo due to the structure of current political processes. Yet it is vital that these interests are represented, for political, economic, moral and legal reasons: not only do concepts of intergenerational equity militate in favour of representation, but it is also more likely that these decision-making processes will reach an optimal result (even for present generations) when these interests are properly weighed.

Past scholarship has noted the need to include explicit pathways in governance structures for accountability to the rights and needs of future generations, but, as yet, such pathways have not yet been fully explored in relation to existential risk. Where such pathways have been adopted, they usually relate to environmental risks like climate change. Our paper hopes to fill this gap in the existential risk literature.
We examine how future generations are currently represented in international and domestic governance systems. We trace philosophical and legal arguments for concerning ourselves with the rights and representation of future generations – why should we care about future generations, and why are they not currently adequately represented? – drawing in part on the intergenerational equity literature that exists in relation to climate change governance. We explore what these rights look like, what 'success' looks like in terms of representation, and how success may be measured. In addition, we explore ways in which the representation of future generations can be improved in the UK, and discuss the relative merits of various options which have been deployed elsewhere. We aim to present a solutions-based approach to the challenge of intergenerational representation.

We submit this paper as part of a joint policy project of The Wilberforce Society and Future of Sentience (FuSe) at the University of Cambridge.

Natalie Jones is a PhD student in Law at Trinity College, Cambridge, studying participation and transparency in international law-making processes. She is a Managing Editor of the *Cambridge International Law Journal*, Co-Chair of Positive Investment Cambridge, and a Commonwealth Scholar. She won the Whewell Scholarship in International Law during her LLM studies at Cambridge. Natalie hails from New Zealand, having graduated in law and in physics from the University of Canterbury. For the past few years she has been a New Zealand youth delegate to the UN climate change negotiations, and follows the multilateral climate regime closely.

Brook Kohn and Paula Brooks

**Building resilient local communities: A Franklin County, Ohio case study**

In local communities, biodiversity, climate, and the environment play key roles. Agriculture, safe drinking water and human health are just a few essential, existential resources that rely on these earth systems. However, given the changing climate, each of these resources faces numerous substantial if not existential risks. In the Midwestern United States, Franklin County, Ohio faces several climate change related threats including: extreme weather events, derechos, flash flooding, polar vortexes, heat related events and illnesses such as strokes, asthma, and other respiratory issues, and food system security disruption.

In response, Franklin County has worked to build community resilience, beginning with a major policy initiative in 2006 that required all policy decisions be made with an eye toward sustainability. This policy was embedded in budget and purchasing decisions as well, in a $1.47 billion annual budget serving 1.2 million residents. Franklin County has followed this policy over the past 10 years. Since then, we have embarked on a plethora of projects, initiatives, and policy decisions that work to mitigate the risks posed by a changing climate, and provide for adaptation.

In 2014, Franklin County Commissioner Paula Brooks joined 24 other American elected government officials on a federal Task Force to survey U.S. communities, seeking solutions to help build resilience across the United States. This work culminated with a report presented to President Barack Obama, who has since acted on the group’s recommendations through the federal agencies and with Executive Orders.
In 2016, Franklin County further acted on the Task Force’s work by establishing a self-replenishing bond fund, which finances energy projects in the community that build energy efficiency and conservation, resilience, and harden infrastructure. This is an example of how a local government can act to mitigate risks posed by a changing climate that is slated to become more impactful of humans over time, especially if left unchecked. We continue to seek more ways to build greater resilience to climate and environmental risks. In this paper we will examine some of the additional policy initiatives we currently see possible in the future. Insightful feedback from the Centre for the Study of Existential Risk will be a major outcome of our conference attendance.

**Paula Brooks** was first elected to the Franklin County, Ohio Board of Commissioners in 2004 and has served the community in that role for 12 years. During that time, Commissioner Brooks has gained a nationally recognised reputation for creating opportunities for job creation and economic growth, investing in early education and healthcare for the county’s most vulnerable, and identifying strategic approaches that result in sensible solutions for all of the residents of Franklin County. In 2013, Commissioner Brooks proudly accepted a White House appointment to President Obama’s bipartisan Task Force on Climate Preparedness and Resilience, where she joined 25 other governors, mayors, county officials and tribal leaders from throughout the nation to seek out and provide climate change preparedness recommendations to the President.

**Brook Kohn** has been with Franklin County as a Policy Director since 2013. He holds an MA in International Relations from the University of St Andrews.

**Karin Kuhlemann**  
*Climate change, catastrophic risks and population policy*  
The global population is currently projected to increase by one billion within the next 15 years or so, and to increase further to 10 billion people in 2056 and 11 billion by 2088. The added growth is equivalent to the entire global population in the mid-1970s. Independently of rising per capita consumption trends, such vast increase in the number of consumers represents a tremendous challenge to efforts to mitigate climate change. In addition, much of the population growth will take place in areas of the world that are expected to be hardest hit by climate change, multiplying the number of lives in harm’s way. The combination of climate change and population growth is anticipated to create humanitarian catastrophes on an unprecedented scale.

Future population growth is not set in stone. Procreative choices, as other individual choices, can be influenced by policy interventions, and relatively small changes to average family size translate into enormous differences in population size over the longer term. The global population in 2100 could be many billions larger or smaller, depending on how far and how fast fertility rates fall. A smaller global population would be much better placed to reduce overall emissions of greenhouse gases and to provide an adequate humanitarian response to environmentally induced mass migrations. It is surprising, then, how little discussion there has been about population policy as a mitigation and adaptive response to the emerging risks of climate change. This silence is almost certainly explained by conceptual and theoretical confusion about the human right to procreate.

The human right to procreate is standardly assumed to rule out most population policies seeking to reduce fertility rates. I demonstrate that this entails giving lexical priority to the right to procreate over all other rights and interests we have, and that this only makes sense if we assume
that the human rights framework prioritises the increase of the number of human beings in existence while accepting unlimited risks to human wellbeing. Such view of human rights is highly implausible, and must be rejected. The right to procreate is not absolute and the duties it imposes must be compossible with the duties imposed by other rights. As an intermediate conclusion, I argue that considerations of justification and compossibility reveal the right to procreate to be inherently a limitable and conditional right.

I then briefly examine the morality of avoidable risk and of risk imposition, with a particular focus on population growth-related risks. I conclude that common assumptions about the permissible timing and nature of population policy interventions are based on fundamental and serious misunderstanding about population dynamics and about the right to procreate.

Karin Kuhlemann is a lawyer specialising in public and regulatory law and a PhD candidate at University College London’s School of Public Policy. Her research interests include human rights theory, population ethics, population policy, human rights law and practice, women’s rights, intergenerational justice, applied ethics, bioethics and animal ethics.

Karin’s political theory doctoral thesis focuses on the interface between the right to procreate and other human rights, with the aim of identifying the legitimate scope for anti-natalist population policies. Karin has degrees in politics, biology and law, and has been involved with population concern activism since 2012.

Matthijs Michiel Maas

*Politics in the age of weapons containment: insights from nuclear weapons policy for the governance of AI development*

This research session focuses on policy options for safe artificial intelligence development, by examining insights from the history of nuclear weapons development and proliferation. It introduces three models from policy analysis (‘Domestic Politics’, ‘Epistemic Communities’, and ‘Normal Accident Theory’) in order to make the following arguments, respectively:

1. Security concerns over a ‘decisive strategic advantage’ may promote arms races, but they do not guarantee it: instead, the history of (horizontal and vertical) nuclear proliferation suggests there are a number of political factors (elite or public) that sway these decisions and can slow or forestall direct proliferation, even in actors that possess the requisite capabilities.

2. ‘Epistemic communities’ of scientists and experts can play a disproportionately large role, directly and indirectly, in promoting technocratic agreements on arms control and disarmament between great powers. Examining the US epistemic community, which laid the foundations for the Anti-Ballistic Missile Treaty, provides lessons regarding the conditions under which expert networks can shape and affect great power cooperation on differential technological development, the strategies they can deploy to do so, and the risks involved in such a project.

3. ‘Normal Accidents’ with unanticipated and uncontrollable failure modes are inherent and inevitable to highly complex, tightly coupled systems and organisations. This imposes severe limits to the safety of any research project or operation working with complex technologies under competitive conditions—and including artificial intelligence—demonstrating the limits to safety.
Accordingly, domestic politics, epistemic communities and normal accidents all play a key role in determining the viability, effectiveness, and possible unintended consequences of possible AI research governance arrangements or strategies aimed at avoiding arms races, facilitating political cooperation, and retaining operational safety. These phenomena illustrate both pitfalls and opportunities for any anticipatory long-term policymaking aimed at ‘maxipok’ AI development, suggesting future directions of research into the causes and cures of competitive arms races, the foundations and catalysts of global cooperation, and the limits to fail-safe project management in the context of technologies such as artificial intelligence.

Matthijs Maas is a Junior Associate at the Global Catastrophic Risk Institute and a researcher with the Global Politics of Artificial Intelligence research group, where his research focuses on the global political dynamics surrounding the militarisation and regulation of AI. He has a background in non-proliferation studies and nuclear deterrence stability, and trend forecasting and regional conflict dynamics. He has previously worked as analyst at Dutch think tanks, The Netherlands Embassy in Beirut, and as web developer. He holds a Masters of Science (distinction) in International Relations from the University of Edinburgh.

Christopher Markou

*Complexity Theory as a paradigm for Artificial Intelligence regulation*

Imagine yourself in one of two worlds. The first is a completely lawless world. There are no police to arrest someone who harms you, no courts to help resolve contractual disputes, and no regulatory oversight for the nuclear reactor that powers your city or the banks you entrust with your savings. The second is a world replete with exhaustive and omnipresent laws and social controls. Free will is non-existent, a cause of action exists for the slightest of inequities, and laws change drastically without warning. If you had to choose, which of the two worlds would you prefer? You would probably yearn for some mechanism to preserve social order and prevent society from devolving into chaos in the lawless world. In the world of omnipresent legal control and non-existent free will you would probably yearn for the ability to live and behave within reasonably defined limits, think and speak freely without fear of censure, and explore choices without being bound to them.

Our world is a point of compromise between these two examples; but where is the balance point between too many and too few laws and regulations? Can we strike a balance between too much and too little free will? Too much and too little justice? Is there a point where individuals are free to make choices, where third-party rights and the regulation of society are optimised in a way that all members are happy, free and prosperous on terms they have discretion to set? If such a point exists, how do we identify it?

More importantly, if we can identify that point, how do we balance all the competing interests and variables that exist in society? These are the fundamental questions posed by the scientific theory of nonlinear dynamical systems and they are the primary subject of this paper.

Specifically, this paper applies insights from the study of nonlinear dynamical systems to the legal and regulatory challenges posed by the rapid development of Artificial Intelligence (AI) and the purposes to which it is being directed. Like nuclear power and genetic research, AI is a classic risk/reward technology. If developed safely it could bring enormous benefits to society.
If developed recklessly and irresponsibly it could pose significant risks. But how should legal systems, courts and regulatory agencies approach managing the potential risks of AI systems without unduly constraining their development? Where is the line between too much and too little control? This paper seeks to open up a new line of thinking for how legal systems should approach the regulation of AI – and by extension other emerging technologies with potentially transformative effects for society – and identify what structural, procedural and moral questions might be better framed using insights from nonlinear dynamical systems theory and social systems theory.

Christopher Markou is a third-year PhD student at the Faculty of Law at The University of Cambridge. His doctoral research, entitled ‘Legal Autonomy and Technological Change: A Systems Theoretical Analysis of Artificial Intelligence’, is supervised by Professor Simon Deakin (Peterhouse) and Professor Ross Anderson (Trinity). Christopher’s research is generously funded by a Doctoral Fellowship provided by the Social Sciences and Humanities Research Council of Canada. He writes on various issues related to technology and society, and his work has appeared in The Guardian, Social Europe and The Huffington Post, where he is a featured writer on law and technology.

Dan Moinard, Matija Franklin, Kartik Vira, Khuzaimah Saeed, Nikolas Bernaola Álvarez, and Beth Barnes, University of Cambridge, The Wilberforce Society, and Future of Sentience Cambridge

Building a framework for accurate forecasting

Mitigating existential risks requires good decisions to be made in anticipation of particular events. These rely on high-quality predictions about what issues will arise, and what the result of different proposed policies or risk management approaches will be. Techniques for knowledge aggregation and elicitation exist that can substantially improve the quality of forecasts, but these are not widely applied.

In general, recent technological advances have helped forecasters across various fields to build increasingly complex models and share huge amounts of data. However, little effort has been put in assessing whether such models and non-mathematical analyses genuinely yield accurate results. Famously, Superforecasting author Philip Tetlock jokingly summarised his study on prediction accuracy by stating “the average expert was roughly as accurate as a dart-throwing chimpanzee”. Comparing this to the importance of decisions relying on subjective judgement from experts, which may have far-reaching consequences, it is clearly time to apply a true scientific method to forecasting.

Vague phrasing and lack of precision, including in the predicted events’ timescale and occurrence probability, are among the main current issues. This results in unnecessary misunderstandings and confusion, and seriously restricts verification and accountability for predictions. The latter are crucial to improve forecasting methods, as well as to maintain the level of understanding and trust required from both the general public and policymakers.

In this paper, we explore where and how the best forecasting techniques could be applied more extensively to decision making. We present the preliminary results of our study into the current research on this topic, the past studies that have radically improved forecasting in certain sectors
including the US intelligence community, and the potential for impact of assessing and improving current forecasting methods and channels. We also suggest an initial set of guidelines that can be applied broadly to help assess and improve forecasting. These range from simple cognitive precautions that each individual can apply to changes in institutions’ attitude towards forecasts and their internal organisations.

In particular, we discuss various specific techniques and areas of application. These include prediction markets as an example of a forecasting tool that could be more widely applied, and artificial intelligence as an illustration of a technology where predictions are hard but crucial to beneficial technological development.

Our goal is to use these suggestions to promote the further study of the implementation of accurate forecasting methods in a wide variety of contexts, and as a first step towards proposing a unified framework of policies implementable within companies and governmental agencies.

**Dan Moinard** is an editor at The Wilberforce Society, the student-run think tank based at Cambridge University. He is also a PhD candidate at the University of Cambridge's Department of Physics; his thesis project focuses on the development of optical characterisation methods for superconducting detectors. His previous education in mathematics and physics highlights his strong interest for the use of mathematical methods to describe both natural and human processes, in particular for forecasting applications.

**Achim Rosemann**

**Gene drive technology: challenges to global risk management and governance**

Gene drive technology is an emerging research field that enables the deliberate shaping of genomes, organisms, and ecosystems. Because it facilitates the spreading of genetic modifications in living organisms at a much faster rate than selective breeding or older forms of genome editing, gene drives allow to directly influence the evolutionary process and to intervene into the function and diversity of life as never before. This situation will generate lasting ecological consequences and raises critical ethical, regulatory and safety issues, which governments, publics and scientists are just beginning to consider. In particular, gene drive technology is associated with numerous environmental and healthcare risks, such as irreversible alterations of ecosystems, bioterrorism and the promotion of dangerous pathogens. There is a widespread consensus that international regulatory frameworks are needed to address these risks. However, reliable forms of ecological and societal risk assessment for this technology field are only emerging and there is disunity among experts on which kinds of evidence are needed to support the release of gene drive applications into the environment.

Another problem is that emerging regulatory and risk mitigation strategies are expensive, technology-intensive and require highly developed regulatory and scientific infrastructures. For low-to-middle-income countries that engage in gene drive research it will be difficult to adopt and implement these emerging international standards. In addition, as social science studies on the agricultural use of GM crops have indicated, divergent economic and political priorities and differing availability of resources are likely to result in variation of implementation and the emerging of informal commercial applications and regulatory grey areas. These factors are likely
to increase shared global risks of gene drive applications, and to constitute a significant challenge to the consistent control of this technology field.

In this presentation I will explore these challenges and their potential implications in greater detail. In part one, I will provide insights into current understandings of (emerging) governance and risk mitigation strategies, and I will examine these strategies in the light of global inequalities. In part two I will examine challenges to public deliberation and public engagement in the gene drive field at a more inclusive, international level. Public engagement has in a recent report of the US National Academies of the Sciences been defined as a crucial precondition and key element of the international governance of gene drive research. However, as I will show, the design, implementation and validity of possible deliberation strategies in the gene drive field is complicated not only by practical and logistic problems, but also by the scientific characteristics of this technology and the geographic dispersion of its possible risks.

**Dr Achim Rosemann** is a research fellow at the Centre for Education Studies, Warwick University. He has a background in STS, Medical Anthropology and East Asian Studies and joined Warwick in 2015 from the University of Sussex where he obtained his PhD in 2014. He is the PI of the Wellcome Trust funded project: "Human germ line editing in global context: challenges and needs from a UK perspective", collaborating with the Nuffield Council of Bioethics. Achim’s research interests concern the social, political, economic and regulatory dimensions of life and health science research, with a particular focus on scientific and social developments in China and East Asia.

**Anders Sandberg and Miles Brundage**, *Future of Humanity Institute, University of Oxford*

**The principle of differential technology development**

If something is technologically possible, it will almost certainly be attempted by someone (the "technology completion conjecture"). Regulation may discourage or shape technology use, but this may not be enough for risky or radically transformative technologies. Regulation is also very limited in its ability to handle future technologies, and banning development of a problematic technology is often unlikely to work while potentially retarding related beneficial technologies.

Another approach is to aim for differential technological development: the order in which technologies arrive matters. If potentially harmful technologies are developed more slowly than other technologies reducing their risks, their benefits will become available with less risk. This can be achieved by focusing funding and research priorities on the harm reduction technologies.

This talk examines the potential and limits of differential technology development, examining case studies from energy, ICT, and biotechnology. In particular, it reviews evidence that it is feasible to speed up protective technology and that protectiveness can be predicted ahead of time. It looks at the relationship of differential technology development to responsible research and innovation as well as the precautionary principle. This is placed in the context of existential risk reduction.

**Anders Sandberg’s** research at the Future of Humanity Institute centres on management of low-probability high-impact risks, societal and ethical issues surrounding human enhancement, estimating the capabilities of future technologies, and very long-range futures. He is currently senior researcher in the FHI-Amelin industry collaboration on systemic risk of risk modelling.
He is research associate to the Oxford Uehiro Centre for Practical Ethics, the Oxford Centre for Neuroethics, the Center for the Study of Bioethics (Belgrade), and the Institute of Future Studies (Stockholm). He is on the advisory boards of a number of organisations and often debates science and ethics in the international media.

Anders has a background in computer science, neuroscience and medical engineering. He obtained his PhD in computational neuroscience from Stockholm University, Sweden, for work on neural network modelling of human memory.

**B. Stahl, S. Rainey, M. Shaw, E. Harris, Centre for Computing and Social Responsibility, De Montfort University, UK**

**Ethics management in the Human Brain Project**

The Human Brain Project (HBP) brings together a number of activities, including animal, human, cognitive and theoretical neuroscience as well as platform development in the fields of neuroinformatics, high performance analytics and computing, medical informatics, neuromorphic computing and neurorobotics.

This combination of activities offers the possibility of ground breaking insights that can substantially change or accelerate the development of artificial intelligence. The exact capabilities of these new technologies are still difficult to predict but their development, seeking to capitalise on our understanding of animal and human brains, leads to high expectations with regard to their impact.

The flipside of these hopes for the development of novel technologies is that they may constitute existential risks that are difficult or even impossible to predict. The current high level of attention to the risks of AI and robotics that is inspiring a flurry of activities on policy, funding and research project levels indicates the degree of concern that these developments raise.

The HBP has been aware of these concerns and has created a set of activities dedicated to questions of ethics and society. These are organised around the principles of Responsible Research and Innovation (RRI). RRI, in the interpretation adopted by the UK Engineering and Physical Research Council (EPSRC), suggests that research needs to include anticipation of future consequences, reflection on the rationale and justification of research, and engagement with various stakeholders, and translate these activities into action. It is assumed that incorporating these principles in all aspects of research will render it responsible. The HBP has implemented these principles in four work packages, which cover foresight, conceptual and philosophical reflection, public engagement and ethics management. If the HBP raises existential risks, then these activities in the ethics and society section should be able to identify them and find appropriate ways of dealing with them.

In our presentation we will focus on the dimension of ethics management. Ethics management interacts with the scientific and research activities of the project in various ways, for example by instituting a research ethics rapporteur programme which reaches out to all other parts of the project; by supporting an external Ethics Advisory Board; by running a reporting system and supervising compliance by researchers with regulation and legislation.
We argue that ethics management is a crucial component of the management of significant risks, as it offers a mechanism that can link the abstract concepts of ethics and risks into organisational practice.

We will conclude the talk by exploring the limitations of this approach. Ethics management is currently confined to the internal processes of the HBP and therefore not in a good position to question the very principles of the project. We will argue that questions of the management of existential risks would benefit from comparable institutions on the funding, policy and international levels, which could be tasked with finding practical ways of guiding research activities in directions that are desirable and acceptable. At present, the UK EPSRC is preparing the provision of RRI services for the ICT community and we will suggest that this may be a step in this direction.

**Bernd Carsten Stahl** is Professor of Critical Research in Technology and Director of the Centre for Computing and Social Responsibility at De Montfort University in the UK. His interests cover philosophical issues arising from the intersections of business, technology, and information. This includes ethical questions of current and emerging of ICTs, critical approaches to information systems and issues related to Responsible Research and Innovation. He works on a number of projects in the area of Responsible Research and Innovation, which includes his role as Ethics Director of the Human Brain Project.

**David W. Wood**

**Lessons from 10 years of public meetups addressing existential risk**

Like all emerging disciplines, x-risk management needs to evolve its processes and its character as it moves from the fringe of public discourse towards a larger share of social attention. This paper provides suggestions for what these changes should include.

Specifically, this paper distils learnings from reflections on the successes and failures in communications that have occurred at both public and private meetings over the last 10 years – meetings where the author has been either the chair or a speaker, and where the communications have addressed existential risks and related topics. During this time, the author has grown the largest and probably the most engaged meetup in the world that covers such topics – see [https://www.meetup.com/topics/singularity/](https://www.meetup.com/topics/singularity/) and [https://www.meetup.com/London-Futurists/](https://www.meetup.com/London-Futurists/).

The paper also draws on insights from the author’s 25 years of active involvement as an executive at the heart of the mobile computing and smartphone industries – a career that saw rollercoaster swings between triumph and disaster – and from his more recent experiences as a futurist consultant working with organisations and individuals to improve their readiness for forthcoming disruptive changes.

Overall, the suggestions covered in the paper can be split into eight areas:

1. Learning from business – crossing the chasm from early adopters to the mainstream market
2. Learning from change management – why most change initiatives fail
3. Learning from Hollywood – the power of narrative
4. Learning from futurism – the art of scenario analysis
5. Learning from the philosophy of science – the exploration of paradigm clashes
6. Learning from politics – the importance of building alliances
7. Learning from marketing – identifying and focusing on key influencers
8. Learning from large-scale software development – taking advantage of Agile.

**David Wood** was a pioneer of the smartphone industry, co-founding in 1998 Symbian, the creator of the world’s first successful smartphone operating system. Software written by his teams was included in half a billion smartphones over the following years, from companies such as Nokia, Motorola, Sony Ericsson, Samsung, LG, Fujitsu, and Panasonic. He also spent three years as CTO of Accenture Mobility.

He is now a full-time futurist speaker, analyst, writer, and publisher. He is the author or lead editor of six books, including *Anticipating 2025*, *Smartphone and beyond*, and, most recently, *The Abolition of Aging*. He also heads up London Futurists – a non-profit networking meetup with over 5,000 members – and has chaired over 150 public events on technoprogressive and futurist topics.

He spent eight years at Gonville and Caius College, Cambridge, studying first Mathematics and then the Philosophy of Science. He has an honorary doctorate in science from the University of Westminster, awarded in 2005 in recognition of his services to the smartphone industry.

He is a fellow of the IEET, Secretary of the Board of Directors of Humanity+, and a fellow of the RSA.
Notes