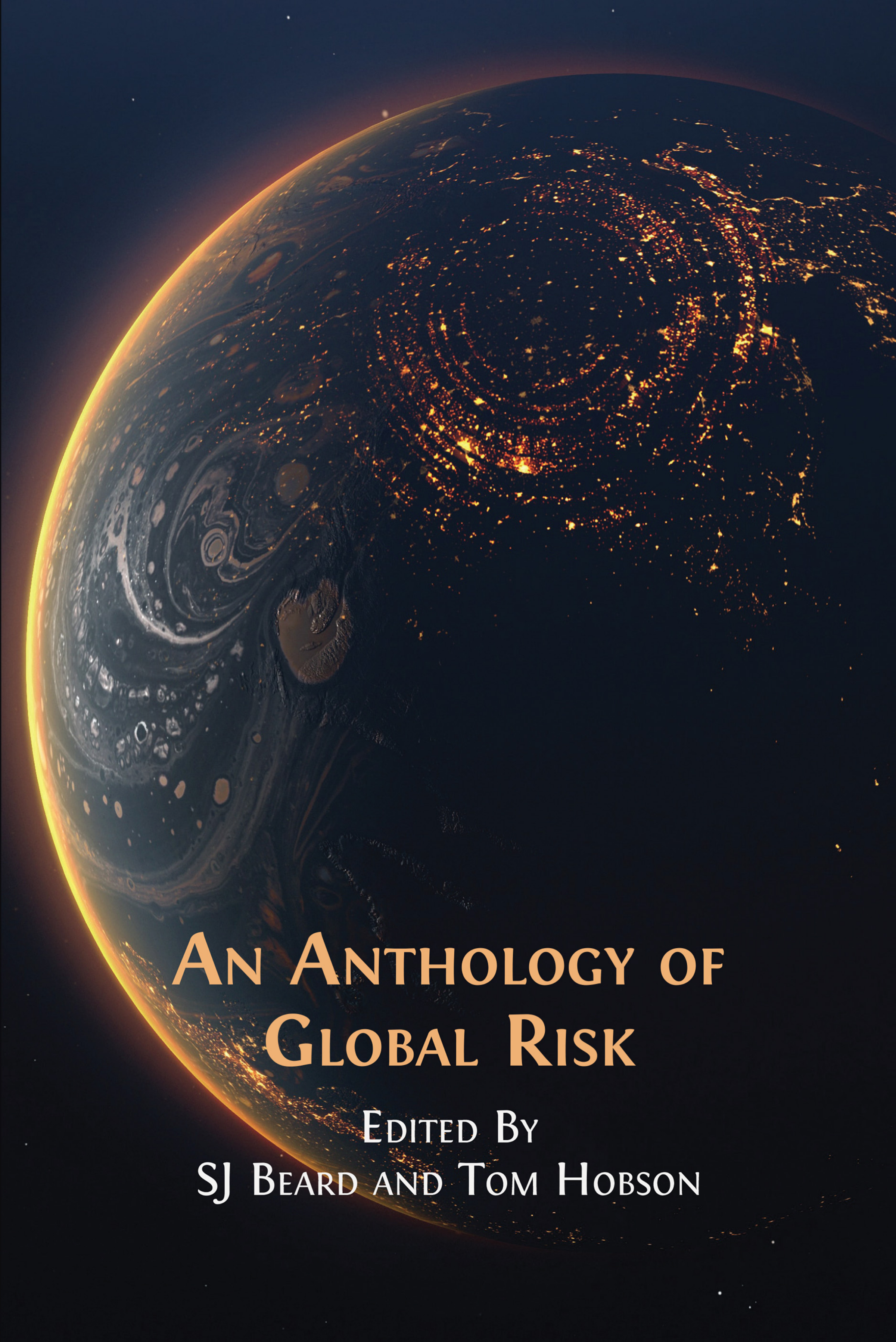


An Anthology of Global Risk

Edited By
SJ Beard and Tom Hobson

AN ANTHOLOGY OF GLOBAL RISK

An Anthology of Global Risk

Edited by SJ Beard and Tom Hobson

OpenBook
Publishers

https://www.openbookpublishers.com

SJ Beard and Tom Hobson (eds), *An Anthology of Global Risk*. Cambridge, UK: Open Book Publishers, 2024, https://doi.org/10.11647/OBP.0360

Digital material and resources associated with this volume are available at
https://doi.org/10.11647/OBP.0360#resources

ISBN Paperback: 978-1-80511-114-6
ISBN Hardback: 978-1-80511-115-3
ISBN Digital (PDF): 978-1-80511-116-0
ISBN Digital eBook (EPUB): 978-1-80511-117-7
ISBN XML: 978-1-80511-119-1
ISBN HTML: 978-1-80511-120-7

DOI: 10.11647/OBP.0360

Cover image: Javier Miranda, Alien planet, June 18, 2022, https://unsplash.com/photos/nc1zsYGkLFA
Cover design: Jeevanjot Kaur Nagpal

Contents

Introduction

> Would that a lion had ravaged mankind; rather than the flood,
> Would that a wolf had ravaged mankind; rather than the flood,
> Would that famine had wasted the world; rather than the flood,
> Would that pestilence had wasted mankind; rather than the flood
>
> — *The Epic of Gilgamesh*, tablet 11

Humans have been living under the shadow of global catastrophe for a very long time. For most of our history, the risk of catastrophe was understood to be supernatural, serving to make it more threatening and fearsome. By some accounts the expected global catastrophe would strike down only the sinful and wicked and bring the blessed to a new life in a better world. By other accounts it may have been seen as part of the natural cycle of life, a cosmic extension of the cycles of birth and death, spring and autumn, rise and fall. Invariably global catastrophes were not the final word for humanity. Although global catastrophes have never been the final word for humanity, always accompanied by promises of salvation and renewal, they nevertheless were maintained as awesome prospects to be feared.

This anthology centres on very different kinds of risk: naturalistic, disastrous, and potentially final calamities. However, that does not mean we should not be concerned with these tales from our past. Old myths about the world's end (from the Christian apocalypse to the Norse-pagan Ragnarök) remain key touchstones for our society, culture, and even politics. Perhaps the most influential of all of these myths is also the oldest and most universal — the deluge. The story of a great flood that was once sent to Earth by an angry god or gods to wipe out humanity is a story told by many cultures around the world. Usually, one human being is forewarned of this impending disaster, however, and is able to escape by building a boat to carry him, his family, and some selection of plants and animals to repopulate the earth. This story

https://doi.org/10.11647/OBP.0360.00

has been found across the Mediterranean Basin and throughout South and Southwest Asia, with comparable stories being told in many other parts of the world. For readers of this anthology, the story might be most familiar by the role it plays in major world religions, including Judaism, Christianity, and Islam (as the story of Noah) and Hinduism (as the story of Vaivasvata Manu). It remains a very popular story for young children all over the world, and is almost certainly the most common introduction most people reading this book will have received to the idea of a global catastrophic event. Yet it is a story that predates all of these religions, with the earliest recorded versions (like the one quoted as part of the *Epic of Gilgamesh* above) dating back over 4,000 years. It may even be the very oldest story to have been passed down to us today.

So, what does this ancient myth tell us about global catastrophes? We are told that the world was nearly destroyed by a single disastrous event (the flood), caused by an exogenous force (the gods), but which happened as a direct result of the faults and failings of humanity. In the story of the great flood, humanity survived because one individual was granted foreknowledge of this catastrophe and was able to take action to save themselves as well as a sufficient number of people, plants, and animals to repopulate the world.

It may be that the story originated in experiences with catastrophic flooding in the river valleys where early civilisations tended to form (and that in places where such flooding was less common, such as in Eastern Iran, the story would survive but with a different agent of disaster, such as a hard winter). It might also be the case that the story reflects the religious sensibilities of the age, in which centralised religion demanded increasingly strict adherence to its laws and requirements on pain of divine punishment. The fact that a story can be so widespread, culturally established in our oral tradition through decades of retelling, suggest that — regardless of origin — certain elements make for a narrative that can withstand the test of time.

It is for this reason that the anthology begins with a discussion of the flood myth. Clearly a compelling story, its elements have been reproduced time after time when we come to think about the end of the world, from speculation about an AI that, due to the imprudent haste with which it was developed, is indifferent to human values and thus chooses to eliminate us, to the hope that we might survive a nuclear or volcanic winter in a bunker or on an island refuge; from a tendency to talk about climate change as an exogenous force that is punishing humanity for our misdeeds, to a desire to predict exactly what kind of biological catastrophe is most likely to bring about a global catastrophe. Our vision of extreme global risks in the early 21st century seems to eerily mirror the stories of our ancestors, even when translated through our present-day claims to rationality and objectivity.

Stories serve to pass down knowledge, ideas, and judgements about how the world is and what it might become; indeed, as Chapters 1 and 8 of this volume describe, they have played, and continue to play, important roles in the development of Existential Risk Studies. However, they also serve as sense-making tools, providing ways to interpret the world around us, its immutability or transience, and the futures we might aspire to or fear. The ability to tell stories, or at least the ability to propagate them and have others listen, is also bound up in social relations and takes place within the material contexts of a given historical moment. Stories do not emerge, fully formed, into the world. Stories are told, heard, retold. Their narratives are reshaped and their endings reimagined. The evolution of the flood narrative over time should also be understood as being shaped by the social relations from which each successive iteration emerges.

To be clear, this does not mean that the resulting ideas are misguided or misinformed. However, it does mean that we should approach them with due care, knowing that we ourselves have been shaped by ancient myths which give meaning and power to certain world perspectives. It is quite possible to tell very different stories about global catastrophes: stories in which humanity is damaged by long-term, slow-moving processes that are endogenous factors in our socio-technological systems, arising from blameless aspects of human nature, or stories in which survival is achieved via a broad awareness of many possible disasters, causing us to increase resilience for all of humanity. It is just that these are not such good stories, and they are never going to capture people's attention in the same way.

The chapters in this volume all contribute to the development of a truly secular approach to extreme global risk, in that they show how we can make significant advances in understanding and managing risk, as well as how we can challenge traditional catastrophe narratives, and create new ones to fit the evidence we are gathering.

To begin with, we can broaden the ways in which we think about extreme global risk. Chapter 1, *Ripples on the Great Sea of Life*, examines the history of how our understanding of this risk has developed over time. The first naturalistic accounts of human extinction and other global catastrophes came from artists and speculative fiction authors looking for new and interesting stories to tell about the end of the world. However, as science and technology developed rapidly through the 19th and 20th centuries, an increasing range of scientists expressed concern that this was a real possibility coming our way. What drew these diverse concerns together, at the dawn of the 21st century, was initially a group of transhumanists who feared that uncontrolled advances in artificial intelligence threatened not only the realisation of their own vision of technological utopia, but also the very survival of humanity. This prompted the establishment of an interdisciplinary community of researchers who saw their goal as charting a safe passage through this "time of terrors" without triggering an existential catastrophe whilst still advocating for further research into artificial intelligence and other technologies so that they might reach the end state they desired. This initial group has been enlarged and diversified by subsequent events, most notably the emergence of the Effective Altruist movement as a substantial source of both additional resources and researchers, and the entrance of, and engagement with, researchers from outside of this community who agreed that extreme global risk was an important problem, if not for the same reasons. The legacy of this history can still be seen in many aspects of the field. Existential risk research still maintains an (arguably disproportionately) strong focus on hazards that could emerge from technologies that many people in the field also see as highly worth developing like AI and biotech; and much of the research in the field remains guided by a common set of ethical and epistemological commitments underpinned by ethical consequentialism and Bayesian epistemology, even though these are not directly related to existential risk.

The remaining chapters in Section 1 all grapple with, build on, and challenge this legacy in a variety of ways. A common theme among these chapters is the need to move away from the most direct and straightforward kinds of existential catastrophe (the naturalistic equivalents of Noah's flood) and towards complex risk assessment that considers a far wider range of possibilities and factors. Chapter 2, *Democratising Risk*, offers a critique of the original paradigm of Existential Risk Studies, what it refers to as the Techno-Utopian Approach. It argues that it is elitist and methodologically limited, so should be replaced with new, more participatory and democratic ways of thinking, which focus instead on complex risk assessment and are transparent about their commitments. Chapter 3, *Classifying Global Catastrophic Risk Scenarios*, provides a framework that helps to meet some of these goals by understanding global catastrophe scenarios from a systemic perspective, moving away from individual scenarios in order to consider convergent risk factors, including systemic interdependence and mitigation fragilities. Chapter 4, *Governing Boring Apocalypses*, provides a complementary framework that rejects a hazard-centric approach to risk, and moves towards considering vulnerabilities (i.e. aspects of humanity and the systems we rely on that make us susceptible to being harmed by hazards) and exposures (i.e. the ways in which hazards and vulnerabilities come into connection with one another); not merely to better understand the full nature of risk, but also because these often provide additional mitigation opportunities. Finally, Chapter 5, *Existential Risk, Creativity, and Well-Adapted Science*, asks fundamental questions about what kind of science is best suited to studying extreme global risk. The chapter makes a strong case that Existential Risk Studies needs to be creative, in the sense of exploring a wide range of hypotheses, rather than seeking to exploit a smaller range of more likely hypotheses, and that as a field it operates within incentive structures that tend to push science towards being more conservative. Countering this in order to achieve the kind of science that is best adapted to its purpose requires exactly the kind of reflexive work that the chapters in this section, and elsewhere in the volume, set out to provide.

Section 2 turns from broad questions about the nature of Existential Risk Studies as a field to consider the methodologies, tools, and approaches for studying it. Some key themes from these chapters include a focus on the value of rigorously implementing methodologies rather than jumping to judgement, even if this is well informed, and the importance of making use of foresight tools that explore a wide range of possible futures rather than trying to forecast the most likely or dangerous among these. As existential risk researchers we have found that one of the most common questions we get asked is 'what should we be most worried about?', following the Noah narrative that successfully surviving a global catastrophe requires us to predict exactly what it is going to be. However, these methodologies provide far more expansive and inclusive ways of studying extreme global risk that avoid this way of thinking entirely.

The first three chapters in this section survey a wide range of different methodologies that can be applied within Existential Risk Studies. Chapter 6, *An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards*, provides a wide-ranging survey and evaluation of methods that have been used for the quantification of risk. It argues that there is no perfect methodology in the field but that it could benefit from a greater degree of methodological pluralism. More importantly, however, the chapter also argues that methodologies need to be applied more transparently and rigorously in order for researchers to engage critically with the limits and interpretations of whatever methods they are using. Chapter 7, *Scanning Horizons in Research, Policy, and Practice*, provides a more focused survey of horizon-scanning techniques. These are structured expert elicitation techniques that both combine information from diverse communities of practice and allow these same communities to sort, verify, and analyse this information to produce better collective judgements, generally aiming at identifying emerging threats, issues, and questions for further research. Chapter 8, *Exploring Artificial Intelligence Futures*, focuses on different methods, that are accessible to researchers from the humanities, for exploring futures of AI. These range from engaging with science fiction and the work of individual disciplines such as philosophy, economics, and risk analysis to participatory methods for bringing diverse groups together. The chapter argues that there is significant potential for more work to be done on the formation and use of participatory role-play scenario tools in particular.

The final three chapters in this section turn to describing three specific methodological tools that have been developed or improved by scholars in this field to better study extreme global risk. Chapter 9, *Accumulating Evidence Using Crowdsourcing and Machine Learning*, describes the creation of TERRA, a semi-automated literature review tool designed to expand the evidence base for Existential Risk Studies. Chapter 10, *The Mortality of States* (*MOROS*) *Dataset*, provides an example of using historical data about the lifespan of political states, MOROS, to study societal collapse and to better understand political institutions that are highly relevant for extreme global risk. Finally, Chapter 11, *Enabling the Participatory Exploration of Alternative Futures*, discusses the ParEvo technique, which enables groups to participate in the construction, exploration, and evaluation of divergent narratives about different possible futures using an evolutionary process.

Section 3 provides examples of how these developments in Existential Risk Studies have been used to produce new insights about the causes and consequences of extreme global risk. The chapters provide insights on a range of risk drivers, from volcanoes to AI. However, it is suggested that these should not be understood as exogenous factors that are out there trying to get us, but simply as the result of processes we are currently struggling to understand.

For instance, Chapter 12, *Global Catastrophic Risk From Low Magnitude Volcanic Eruptions*, argues that traditional accounts of Global Catastrophic Volcanic Risk focus too much on the explosivity of potential future eruptions. However, the relationship between the size of an eruption and the amount of damage caused is neither straightforward or linear. By plotting active volcanoes alongside critical global infrastructure such as manufacturing and transportation pinch points, the chapter shows how we are especially vulnerable to volcanic eruptions in particular localities, due to the placement of key infrastructure in areas where eruptions could easily damage or disrupt it. Hence, it is possible for even a relatively low explosivity volcanic eruption from the right/wrong volcano to cause harm at the global scale. Chapter 13, *Re-Framing the Threat of Global Warming*, looks at the risk from climate change and provides an empirical evidence base for studying how this could be mediated through food insecurity and societal collapse. By conducting an extensive literature review, the chapter constructs an empirical causal loop diagram that describes the systemic cascades that could be triggered by future climate change, and that are created by the ways we have designed national and international institutions and systems around current climatic expectations. Chapter 14, *Existential Change*, builds on this with a theoretical exploration of what Existential Risk Studies can learn from climate change more broadly, highlighting how the tendency to ask questions such as 'is climate change an existential risk?' misunderstands the nature of risk and fails to learn lessons from other researchers, who have studied climate change, about its likely effects. As a result, we need to move away from thinking about "climate change" as a single force and towards thinking through a diversity of different climate scenarios. Chapter 15, *A Fate Worse Than Warming?*, turns to consider one of the elements of future climate scenarios: the potential use of Stratospheric Aerosol Injection, a technology that injects sulphates into the upper atmosphere, deflecting sunlight and providing a global cooling effect. This has been touted as a possible means of mitigating risks from climate change but it is a risky technology in its own right, and the chapter assesses what these risks are, what we know about them, and how they might be weighed against the potential benefits. The chapter concludes that it is unlikely that we can conclusively ever say whether the Stratospheric Aerosol Injection is "good" or "bad" as so much will depend upon other features of any scenarios in which it is deployed. Chapter 16, *Bioengineering Horizon Scan 2020*, uses horizon-scanning techniques like those described in Chapter 7, to look at emerging issues in bioengineering. It identifies 20 issues including technological, societal, and governance changes that could emerge over a range of time spans and are of highest priority for further research. Finally, Chapter 17, *Artificial Canaries*, shows how we can combine a variety of methods to identify early warning signs (or "canaries") that Artificial Intelligence may be on the brink of increased transformative potential. These take account of both what experts currently know about the possibilities for future AI and where there is currently most uncertainty, so that the warning signs give sufficient room for anticipatory governance frameworks to be put in place to manage this transformation for good rather than ill. By focusing on a broad concept of what transformative AI might be like and how we can learn more about what kind of future trajectory we might be on, this chapter once again highlights the importance of exploring different possible futures and understanding how we continue to shape these through present and future choices.

Finally, Section 4 considers mechanisms for reducing the level of extreme global risk by improved policy-making. These chapters are grouped according to their shared concerns with shaping policies, institutional behaviours, and governance priorities. They take a variety of approaches to undertaking this task, emphasising dialogue, collaboration, equity in representation, and the importance of linking policy-making to scientific expertise. However, they are all clearly focused on prevention, rather than survival, and on spreading power to more people who can use it to collectively achieve common goals, not prioritising the interests of elite latter-day Noahs. The chapters also pose important questions about how we might go beyond reactive engagements with risks from hazards that are considered as already imminent or intrinsic, to instead proactively fostering social, political, and economic conditions more amenable to human and planetary survival.

Chapter 18, *Pathways to Linking Science and Policy for Global Risk*, proposes that engagement with policy-making is a necessary and core component of existential risk research, as an action-oriented discipline. The chapter provides an overview of some of the policy shaping work undertaken by researchers at CSER and highlights promising approaches that scholars might take in the future. Chapter 19, *The Cartography of Global Catastrophic Governance*, takes a more macro-level approach to charting the efficacy and concentration of different GCR governance efforts, proposing a typology that allows for comparison based on risk focus, institutional arrangement, and effectiveness of implementation, highlighting the gaps scholars are best positioned to fill. This chapter provides a map of governance efforts for different GCRs at the time of writing, whilst additionally presenting an analysis of what kinds of action might serve to best increase resilience to GCRs, even in the face of complexity and uncertainty. The remaining chapters focus on more specific contexts, ranging from national policy and institutional design to international diplomacy and private sector investment. Chapter 20, *The Stepping Stones Approach to Nuclear Disarmament Diplomacy*, marks another change of focus and level of analysis as the author provides a reflective account of efforts to build dialogue, and embraces the potentialities for radical change that might be catalysed by even modest incremental improvements in diplomatic relations towards nuclear disarmament. Chapter 21, *It Takes a Village: The Shared Responsibility of Raising an Autonomous Weapon*, considers a specific policy area, defence policy and military procurement towards Lethal Autonomous Weapons Systems (LAWS). It explores how this can be improved by simulating an inquiry that might take place following a LAWS-initiated fatality, and uses the results to show how narrow policies shaped by restrictive notions of "human control" are likely to be insufficient to govern these systems. In Chapter 22, *Representation of Future Generations in United Kingdom Policy-Making*, the focus shifts towards representation and we are prompted to consider how, and why, we might seek to ensure equitable representation of future generations in the national policy-making processes of today. The final chapter of this volume, *Financing Our Final Hour*, provides readers with an empirically grounded analysis of how different modes of pressure and advocacy can influence institutional investors to take seriously the responsibility they have to people and the planet to reduce the re-production of catastrophic hazards in their investment practices.

These chapters also serve to further emphasise the point that extreme global risks, and the means of reducing or preventing them, are never *ex machina*. Rather, they are shaped through an ongoing process of interactions: interpersonal, international, and technological relationships come to the fore in the sections' analyses of how researchers might shape policy and practice in this field. For many of us, these processes can seem very remote, and it is important not to forget how concentrated much of the power over risky scientific, technological, and economic development really is. However, these chapters prove that many of us are already enmeshed within institutions, from parliaments to pension funds, that have the power to influence them. These chapters also promote us to think positively about what better institutions and policies might look like. For instance, Chapter 20's stepping-stones approach starts by drawing on radical visions of how security could be achieved without weapons of mass destruction, while Chapter 23 makes a strong case that large institutional investors, known as universal owners, have strong ethical, legal, and financial reasons to reconceptualise themselves as responsible stewards of the entire economy, and should use their power for collective goods like the reduction of global risk.

Together, the prospect of distributed power and responsibility, and reflection on positive possibilities for how existing institutions could be used in the service of creating positive futures, opens the door to very different ways of thinking about humanity's 21st-century predicament. We are not only living in an age of extreme global risk, including existential risk to the future of humanity, but also living with the possibility of existential hope. That humanity may be heading for extinction is of very limited interest in the cosmic scheme of things, as extinction is ultimately the fate of all species. However, if we rise to the challenges of our age then it is possible that humanity may be the first species in the long history of our planet to have created the conditions for our own extinction and then chosen to do something else. That seems like a project worth pursuing. As Martin Luther King Jr famously put it in his final speech, the night before his assassination in 1968:

> And another reason that I'm happy to live in this period is that we have been forced to a point where we're going to have to grapple with the problems that men have been trying to grapple with through history, but the demands didn't force them to do it. Survival demands that we grapple with them.[1]

We do not wish to claim that Existential Risk Studies has yet earned the right to say that we are delivering Dr King's dream of a world in which people truly face up to the reality of such problems. However, we do share his view that one can be happy to live in a time when the ancient fears of global catastrophes may finally be leading us to at least think about how this work may be done.

Our contention is that this anthology signals something special, the establishment of an entirely new field of study, Existential Risk Studies, and we hope that the chapters within it, and the conversations between them, show how this field is developing. The chapters engage with an issue that is of great concern to many and examine its meaning and foundations, developing methodologies to study it responsibly, revealing new insights about its nature and impacts, and advocating for meaningful change to make it less concerning. They show how we are moving away from the speculative and alarmist and towards the proactive, rigorous, transparent, and accountable. In doing so, it is suggested by the authors that we can move away from the deep myths that have defined our past, towards a creative and engaged science that can help us build a better future.

Notes and References

[1] King Jr, M. L. 'I see the promised land', in J. M. Washington (ed.), *A Testament of Hope: The Essential Writings and Speeches of Martin Luther King, Jr*. HarperOne (1986), pp. 279–86.

I. HISTORY, CONCEPTS, AND NORMS

The chapters in this volume all take as their subject the study of extreme global risks, in most cases extreme to the point of involving either a global or existential catastrophe. As the first chapter in this section, *Ripples on the Great Sea of Life: A Brief History of Existential Risk Studies* by Beard and Torres, suggests, this is not a particularly new research area. Existential risk has been a subject of speculation and research going back at least into the 19th century. However, the notion of a transdisciplinary field of Existential Risk Studies, and the growth of a research community dedicated to it, are far more recent developments. The chapters in this volume emerge from this nascent area of research, and represent many of the key debates in the field. Indeed, this contributions to this volume show, perhaps above anything else, that the key concepts and norms of this community are still very much points under discussion.

To open this volume, therefore, we offer five chapters that provide a range of sympathetic perspectives on this emerging field and how it is developing. It is possible to see these chapters as providing the basis for their own paradigm of existential risk research and/or to see them as substantially critiquing certain views within the field. There are certainly points on which they all agree; such as the importance of systemic risk and looking beyond the most direct and explosive forms of catastrophe, the need for pluralism and interdisciplinarity, and the importance of integrating risk assessment and mitigation in research. However, we hope that by bringing them together the points of divergence and discussion between these chapters becomes clearer, so that the reader can more easily appreciate that they are united not be a shared idea of what existential risk is and how to reduce it, but by a shared commitment to more expansive, open, and reflexive ways of working to answer that question.

The first chapter of this section provides a historical overview of the study of existential risk. It presents an account of the ways existential risk was thought about during the 19th, 20th, and 21st centuries and proposes a possible genealogy of the emerging field of Existential Risk Studies. This chapter emerged from a desire by the authors, one of whom is co-editor of this volume, to see a history of this field written, not simply as an intellectual exercise but also a way of helping the field develop and improve. The authors envisioned their history as making three main contributions: 1) a way of introducing new researchers to the breadth of ideas within the field, and where they had come from, 2) an opportunity to learn lessons from how this work has been done in the past that could be applied to doing it better in the future, and 3) to inspire people to think about how the field might be developing and what its possible future trajectories could be. That the work was successful in fulfilling these objectives is perhaps best demonstrated by the scale and scope of further discussions it has been taken as a starting point for. Scholars — including Thomas Moynihan,[1] Daniel Zimmer,[2] Apolline Taillandier,[3] and Matthew Connelly, among others — have all made contributions to an evolving dialogue, painting different pictures of the trajectory of this field, allowing for greater learning and stimulating more ideas about where it may be heading.

In this particular history the trajectory of the field is told in terms of three successive "waves" of development. The first of these, which occurred largely during the 2000s, saw a small group of researchers conceptualise the idea of existential risk from a specifically transhumanist perspective. These researchers, such as Nick Bostrom and Eliazer Yudkowsky, were committed to bringing about a transition to a post human state that they felt would be better than our present condition but also argued that, along the way, the technologies we were developing were creating risks with the potential to threaten both the lives of presently existing human beings and the potentiality of this "ideal" future. The field of Existential Risk Studies can be seen, in many ways, as these researchers both seeking to produce a coherent research agenda for understanding and mitigating these specific threats as well as to place their work in dialogue with that of many concerned scientists and technology developers who had long worried about humanity's potential to wipe ourselves out. A second wave emerged in the late 2000s and early 2010s, when these researchers sought to increase the power and impact of their field by engaging with broader communities while, at the same time, the burgeoning Effective Altruism movement, whose aim was to identify how individuals could produce the greatest quantity of value with their actions, started to wonder about existential risk mitigation as a priority cause. These two movements saw the field expand and diversify rapidly and lead to ethics, and in particular the forms of consequentialist ethics most common among Effective Altruists, becoming deeply embedded in the methods and concepts of the field. Finally, a third wave has emerged since the late 2010s, in which this expanded field has been influenced by the diverse perspectives of new researchers, many of whom are not interested in, or are even actively hostile to, transhumanism and/or consequentialist ethics and who have also introduced new methods, approaches, and perspectives, many of which relate to systems thinking and complexity.

The second chapter, *Democratising Risk* by Crèmer and Kemp, considers some of the same issues but moves from a historical analysis to a reflection on what the present situation of Existential Risk Studies implies about the field's strengths and weaknesses. In particular, the chapter critically assesses the legacy of pioneering work in the field that shared transhumanist, utilitarian, and longtermist assumptions. The authors characterise this as the Techno-Utopian Approach. One of the implications this has had for the field is that standard definitions of existential risk are concerned less with death or harm suffered by people in a global catastrophe, but rather more with the significant loss of "value" (according to a particular set of assumptions about what is valuable) that this would represent, not merely in terms of actual harm but especially quantity of potential future value in future lives that would never come into existence as a result. The authors argue that a definition like this is problematic because it will invariably tie our understanding of existential risk to the value system of a particular group of people and that this is both philosophically tenuous — for a field that claims to speak on behalf of humanity as a whole — and practically dangerous, in that it carries the potential of justifying the values of this group being imposed on everyone else. They thus argue that the field should separate work on existential and extinction ethics (what is good or bad in relation to existential risk and the future of humanity, how to prioritise work in this area, how to make decisions under uncertainty, and other related questions) and the study of human extinction and other global catastrophes. They also point to a range of other methodological weaknesses in the Techno-Utopian Approach, including a techno-deterministic view of the future and a focus on simplistic, threat based, models of risk assessment and mitigation.

Furthermore, they argue that these features of the field carry the potential to generate negative consequences and "response risks" if its recommendations were to disproportionately influence future decision makers. As a result, they call for a diversified and democratised field that is open to a wider range of assumptions and approaches and seeks to both listen to, and engage with, a far wider community of stakeholders.

While both these chapters present a somewhat critical engagement with the history and possible future trajectories of the field of Existential Risk Studies, there are a number of points of divergence between their analyses. For instance, while the first views Existential Risk Studies as presently evolving through a process of diversification and maturation, the second argues that there are important forces currently seeking to stifle this and ensure that the field does not deviate too far from the ethical frameworks it was founded around. Similarly, the first chapter argues that certain individuals, as well as the field as a whole, can be seen as transitioning between the various waves of existential risk research (indeed, in some ways, the three waves can be seen as typified by three papers from Nick Bostrom in particular — *Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards*,[4] *Existential Risk Prevention as Global Priority*,[5] and *The Vulnerable World Hypothesis*[6] — although each wave, and especially the third, go far beyond the work of this one prominent scholar), the second views different perspectives within Existential Risk Studies as far more bound up with the interests of competing groups and argues for the necessity of democracy and justice to rectify this. These, and other differences, should definitely not been seen as requiring there to be a right or wrong answer about the nature of the field; however, they point to different understanding of it, and we invite readers to reflect on this and draw their own conclusions.

The remaining three chapters in this section move from a consideration of the field of Existential Risk Studies to thinking about existential (and other forms of extreme global) risk and how to study and prevent them. Chapter 3, *Classifying Global Catastrophic Risks*, by Shahar Avin and a team of (then current) CSER researchers, sought to provide an analytical framework for thinking about different kinds of global catastrophe scenario. Rather than drawing upon people's immediate thoughts about what kinds of mechanism could bring about a global catastrophe, it approaches the subject by thinking about the nature of such catastrophes themselves. It shows that in all cases a global catastrophe involves at least one critical system on which humanity depends being pushed beyond the safe limits for supporting our survival and flourishing, one or more spread mechanisms that would cause this effect to be experienced globally (and or prevent it being contained locally) and, crucially, the fragilities in human decision making that caused us to fail to prevent this from happening. This framework serves as both a methodology for thinking clearly about how particular global catastrophe scenarios, from pandemics or volcanic eruptions to environmental or technological catastrophes, might play out, whilst also understanding all of the factors that might trigger such a catastrophe, and hence the avenues we might have for preventing it. This chapter has been foundational to much of the contemporary thinking about the systemic components of Global Catastrophic and existential risk, at the CSER and elsewhere.[7]

Chapter 4, *Governing Boring Apocalypses* by Hin-Yan Liu, Kristian Cedervall Lauta and Matthijs Michiel Maas, provides a complementary analytical framework for thinking about the drivers of risk (rather than the catastrophes that might emerge from it). The chapter argues that when people think about the causes of Global Catastrophic Risk they often focus only on the "hazard" or "threat" that precipitated the catastrophe, such as an asteroid, pathogen, or unaligned Artificial Intelligence. While hazards are important however, decades of work in disaster studies has shown that they are not the only drivers of risk. For a catastrophe to occur, a hazard needs to be combined with two other features: a vulnerability (a factor that makes humanities subject to the harm this hazard might cause) and an exposure (the medium by which the hazard and vulnerability meet). For instance, earthquakes, a quintessential hazard, seldom kill people on their own. The harm that earthquakes do comes about because we create a vulnerability by building and living around structures that an earthquake might cause to collapse upon us, and also because we expose ourselves by doing this in areas where earthquakes happen.[8] The chapter goes on to provide a complete classification of existential vulnerabilities and exposures that, once again, both helps to explore the nature of risk drivers that are already being examined within the field, as well as drawing attention to the possibility of discovering new risk drivers that have not yet been considered. However, perhaps even more importantly, the authors point out that a framework like this not only helps us think about new ways in which things might go wrong but also, by extension, to understand the full suite of tools at our disposal for preventing them. In particular, they argue that once a full inventory of vulnerabilities and exposures has been produced, we can appreciate that technological solutions are not only far from the sole options at our disposal for preventing existential risk but that they might also be harmful. For instance, they can feed into cultural vulnerabilities and exposures, reduce societies' resilience, and also risk breeding a false sense of security; we take comfort in a few easy solutions (such as rapid vaccine development) and this causes us to lose sight of more complex problems (such as inadequate provision of public health).

It is notable that these two chapters first appeared at the same time, and they may appear to be offering the same kind of output, an analytical framework for thinking more expansively about existential and Global Catastrophic Risk. However, they are not the same, nor are they in competition with one another. For many years, people in Existential Risk Studies have talked about existential risks, labelling things such as AI, biorisk, climate change, and nuclear war. However, while these things might be (usefully at times) labelled and understood as "risks", they are *also* technologies, processes, trends, and events. These two chapters do not seek to classify different risks but global catastrophe scenarios (the catastrophic events themselves) and the drivers of risk (the processes and phenomena that come together to precipitate events such as this). It is important to note that these two things cannot be matched up on a one-to-one basis; for instance in Classifying Global Catastrophic Risk Scenarios the authors note that the global catastrophe of a darkening of Earth's atmosphere leading to mass starvation could equally be caused by a range of mechanisms, from an asteroid impact to a nuclear war. Similarly, Governing Boring Apocalypses highlights how the same kinds of vulnerability and exposure (such as just in time food delivery or short-term political decision making) can be involved in precipitating many kinds of catastrophe. We therefore need to apply both of these frameworks in order to fully understand both the drivers of risk and catastrophe scenarios, but also to understand that in doing so we are assessing only one, more or less unified, phenomena of existential and Global Catastrophic Risk, not many distinct existential and Global Catastrophic Risks. This is a conceptual innovation that is still developing within the community.

The final chapter in this section, *Existential Risk, Creativity and Well-Adapted Science*, by Adrian Currie, looks at the kind of science that would be best suited to studying existential and Global Catastrophic Risk. The chapter considers many of the questions raised by previous chapters, such as the value-laden nature of science and the difficulty of studying the interaction of complex systems, and argues that for a field such as Existential Risk Studies to be well-adapted to this situation, it is important for researchers to show a high degree of creativity, raising, exploring, and testing many different kinds of solution, rather than the conservative strategy of trying to identify only those solutions most likely to succeed and not exploring more widely. In this aim, however, the chapter notes that the field runs up against many different features of science that currently work against precisely this kind of creativity. Achieving more creativity requires the field to be multi-disciplinary, pluralistic, and opportunistic, but most of all it requires researchers within the field to identify the sources of maladaptation and ask which of these we might do something about. This means that we need a creative engagement with both existing norms and practices in the field, and also with those taken as best practice within science more broadly, from competitive peer-reviewed funding to the institutionalisation of scientific research. We cannot safely assume that the norms and incentives we are developing will provide the field with the creativity it needs to succeed at its aims.

Clearly, this chapter points us right back to the beginning of this section, and the need to interrogate what the field is and how it came to be this way. However, we hope that it also provides an opening into thinking about the remaining chapters of this book, in which we turn from engaging with the field to thinking more about the problems it aims to solve. In the next section we will look at a variety of methods, tools, and approaches that have been developed by researchers in this community to study existential and Global Catastrophic Risk. In their own way, all of these seek to promote more creative work while also upholding standards of transparency and rigour. However, in moving forward from the groundwork of this section, we suggest that readers proceed to the next section, and indeed to any other work in this field, with some key reflective questions to hand. For instance:

- How did this method, tool, approach, or idea come to be, what community created it and what assumptions, norms, and values shaped it, are there other versions of the same method, tool, approach, or idea that might reflect different ways of thinking about the same problem?
- How do these ways of thinking about existential and Global Catastrophic Risk fit with particular individuals, institutions, or paradigms and the power that they hold? Do they play a role in concentrating resources (not merely economic but also social, cultural, and epistemic) or distributing them more fairly?
- Does this method, tool, approach, or idea allow for a complete analysis of all aspects of a problem, or does it tend to focus on one element that is most readily observed or easy to study and, if so, what is left out?
- How can this idea be used not merely to understand risks better but also to manage them?
- Is this a helpful way of exploring the maximum possible range of possibilities or ideas? Is it a way of exploiting one subset of possibilities and ideas most fully? If it makes trade-offs between these two, how does this work and who gets to decide?

Questions such as these can easily be dismissed as naval gazing by those who wish to jump straight into thinking about a problem. However, one thing that we take from the chapters of this section is that this is a mistake. While reflection can be difficult, and critique can be even more so, understanding the context in which one works, and how one's work fits with the larger field of Existential Risk Studies, is of vital importance if we are to make sure that the community as a whole is finally making progress towards the long-held desire to ensure the safety of humanity, now and in the future.

Notes and References

[1] Moynihan, Thomas. *X-Risk: How Humanity Discovered Its Own Extinction*. MIT Press (2020).

[2] Zimmer, Daniel. *Essence, Process, System: Human Extinction in Political Thought from Aristotle to the Anthropocene* (in preparation).

[3] Taillandier, Apolline. 'From boundless expansion to existential threat: Transhumanists and posthuman imaginaries', in Sandra Kemp and Jenny Andersson (eds), *Futures*. Oxford University Press (2021).

[4] Bostrom, Nick. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9* (2002).

[5] Bostrom, Nick. 'Existential risk prevention as global priority', *Global Policy*, *4*(1) (2013): 15–31.

[6] Bostrom, Nick. 'The vulnerable world hypothesis', *Global Policy, 10*(4) (2019): 455–76.

[7] However, it is worth noting that the chapter itself builds upon previous work, notably the planetary boundaries framework for thinking about anthropogenic environmental disruption and the Boundary Risk for Humanity in Nature framework developed by Seth Baum and Itsuki Handoh, who brought this concept into dialogue with ideas about Global Catastrophic and existential risk. See Rockström, Johan, Will Steffen, Kevin Noone, Åsa Persson, F. Stuart Chapin III, Eric Lambin, Timothy M. Lenton et al. 'Planetary boundaries: exploring the safe operating space for humanity', *Ecology and Society, 14*(2) (2009); Baum, Seth D. and Itsuki C. Handoh. 'Integrating the planetary boundaries and global catastrophic risk paradigms', *Ecological Economics, 107* (2014): 13–21.

[8] See Hörhager, Elisa and Julius Weitzdörfer. 'From natural hazard to man-made disaster: The protection of disaster victims in China and Japan', *Protecting the Weak in East Asia*. Routledge (2018), pp. 139–65.

1. Ripples on the Great Sea of Life: A Brief History of Existential Risk Studies

SJ Beard and Emile P. Torres

Research highlights:

- While thoughts about naturalistic human extinction can be traced back to the latter 19th century among both speculative artists and concerned scientists, the field of Existential Risk Studies (ERS) only emerged in the last two decades and can be characterised by three distinct "waves" or research paradigms.
- The first was built on an explicitly transhumanist and techno-utopian worldview and the risks associated with it.
- The second grew out of an ethical view known as "longtermism", closely associated with the Effective Altruism movement, and is concerned with creating the most value possible.
- The third emerged from the interface between ERS and other fields that have engaged with existential risk, such as Disaster Studies, Environmental Science and Public Policy.
- In adumbrating the evolution of these paradigms, together with their historical antecedents, the authors offer a critical examination of each and speculate about where the field may be heading in the future.

https://doi.org/10.11647/OBP.0360.01

This chapter sketches the history of Existential Risk Studies up to the year 2020. Chapters 3 and 4 provide some of the key original sources for the shift to global systems thinking described here as the third wave of ERS. The continuing influence of speculative fiction on ERS and wider social perceptions around AI and Existential Risk are discussed in Chapter 8.

> But if some poor story-writing man ventures to figure this sober probability in a tale, not a reviewer in London but will tell him his theme is utterly impossible. And, when the thing happens, one may doubt if even then one will get the recognition one deserves
>
> — H. G. Wells, *The Extinction of Man* (1897)

> A colleague of mine likes to point out that a Fields Medal (the highest honor in mathematics) indicates two things about the recipient: that he was capable of accomplishing something important, and that he didn't. Though harsh, the remark hints at a truth
>
> — Nick Bostrom, *Superintelligence* (2014)

There is an emerging scientific consensus that, due to the multiplicity of risks with the potential to cause global catastrophes, *Homo sapiens* is now in the most perilous moment of its 300,000-year history. We face global challenges of such magnitude that, by comparison, all the setbacks and tragedies of human history are "mere ripples on the surface of the great sea of life".[1] Yet if all that we have ever known are such ripples, how can we understand, let alone stop, the tidal waves that threaten to engulf us?

It is thus hardly surprising that a new field focused on the long-term survival of our species is emerging. This has variously been referred to as "Existential Risk Studies" (ERS), "Existential Risk Research" and "Existential Risk Mitigation". For the present purposes, we will use the acronym "ERS", to fit with related fields such as Futures Studies, Science and Technology Studies and Disaster Studies. The aim of this chapter is to explore the historical development of ERS and, in doing so, to identify points of convergence and divergence between different researchers studying existential risk. We argue that there have been multiple ERS paradigms or "waves", i.e., sets of concepts and practices in the sense of Thomas Kuhn (1922–1996). These can be distinguished according to the following issues: (i) *definitions of* *key terms*, (ii) *motivating values*, (iii) *classificatory systems*, and (iv) *methodologies*.

We break these paradigms into four groups, which we consider in successive sections of this work. The next section deals with the history of thinking about existential risk that preceded the emergence of ERS as a unified field of study in the early 2000s, looking at how the topic has been explored by speculative fiction authors and concerned scientists. Section 2 considers the forces that helped to unify ERS in the first decade of the 21st century, which arose from a specifically transhumanist or techno-utopian world view. Section 3 explores a second paradigm, connected with a significant expansion of both interest in and support for this field, related to the growth of the "Effective Altruism" (EA) movement after 2009 and its promotion of ethical longtermism. Section 4 examines a third paradigm, which has emerged in recent years both within certain centres of ERS research and among the scientists from other fields who are beginning to engage with it, that focuses more on global systems and is comparatively less interested in ethics. Finally, Section 5 offers some speculation about the possible future trajectories of ERS and the developments that will drive them: increased scrutiny and public attention, the growing list of existential threats to humanity, and the diversification of the field.

In breaking down the paradigms of ERS into successive waves we do not claim that these represent cohesive social groups or schools of thought; it is notable that many individual scholars have passed between many of them and would not necessarily identify any strong change of mindset in doing so. Nor do we mean to imply that successive waves have succeeded or replaced each other. However, we do claim that roughly combining the work of scholars into these waves tells an interesting and useful story that helps to illustrate and explain the development of the field of ERS. Even more importantly, we hope that it helps to identify how the seemingly disparate and even contradictory claims of scholars can be understood as offering complementary perspectives on a common problem, and thus that our work will help to ensure that ERS remains a coherent field of study as it continues to diversify.

Section 1: The Prehistory of ERS

People in many cultures throughout history have speculated about the possibility of global catastrophes, up to and including the "apocalypse" or "end of the world". Indeed the first story ever to have been written down may well have been the Mesopotamian "flood myth", which tells of a flood that wiped out all but two humans and is familiar to most in the west through its inclusion in the Bible as the story of Noah.[2] However, such speculation has largely been bound up with religious beliefs and invariably ends with the survival of humanity, either on Earth, in an afterlife or via an eternal cosmic cycle of rebirth.[3] In contrast, the notion of existential risk is both absolute (humanity's extinction or ruination is both total and irreversible) and naturalistic (the fate of humanity is to be brought about in accordance with scientific laws of nature). Concern about this kind of catastrophe has been far less common. Indeed, the very idea of *human extinction* is a recent invention. The four primary reasons[4] for this are that:

1. The scientific community largely rejected the possibility that species could go extinct until the French zoologist Georges Cuvier (1769–1832) demonstrated that elephantine bones unearthed in Siberia and North America belonged to mammoths and mastodons.

2. The belief that an ontological gap separates humans from nature, which was prominent at least until Charles Darwin's *On the Origin of Species*,[5] convinced the scientific community that evolution is a fact about the history of all Earth-originating life, metaphysically integrating humanity into the natural order.

3. Religious eschatologies monopolised thinking about the fate of humanity until the 19th century; it wasn't until the 1960s that the "Age of Atheism" commenced, to borrow a term from Gerhard Ebeling.[6]

4. There was no agreement within the scientific community about the existence of potential kill mechanisms (other than

> the second law of thermodynamics) that could annihilate humanity until the second half of the 20th century.

Yet, over the past two centuries, several historical precedents for the modern field of ERS have emerged, and it is worth considering these before turning to the history of this field.

Speculative fiction

Some of the earliest thinking about human extinction in a naturalistic sense are found among artists in the early 19th century. For example, in works by Lord Byron (1788–1824), the infamous romantic poet and father of computer pioneer Ada Lovelace. Lord Byron is reported to have been interested in comets and concerned that humanity would someday perish as a result of a comet impact, while his 1816 poem "Darkness" imagines a future in which Earth becomes lifeless (probably inspired by the after-effects of the 1815 eruption of Mount Tambora).[7] Mary Shelley (1797–1851), Byron's friend and the founder of science fiction, published *The Last Man* in 1826.[8] This tells the story of Lionel, who witnesses the death of all other human beings in the last few decades of the 21st century from a series of apocalyptic events, most notably a worldwide plague, and must come to terms with the fate of the world. Shelley was likely influenced by the loss of her husband (Percy) and many friends, including Lord Byron, in the preceding years. However, she may also have been influenced by the work of her parents, William Godwin and Mary Wollstonecraft, who envisioned utopian futures of social equality and progress, which Mary's own life had often failed to realise. Shelley's novel was not the first of its kind, though, and indeed it was part of a literary genre concerning the fate of "the last man", originating with the 1805 publication of an identically titled work by Jean-Baptiste Cousin de Grainville,[9] which described a future in which the human population dwindles because of infertility.

The discovery of the second law of thermodynamics in the early 1850s inspired new thoughts about human extinction among both science fiction writers and working scientists. For example, in his 1870 book *Sketches of Creation*,[10] the American geologist Alexander Winchell describes an "awful catastrophe which must ensue when the last man shall gaze upon the frozen Earth, when the planets, one after another, shall tumble, as charred ruins, into the sun, when the suns themselves shall be piled together into a cold and lifeless mass, as exhausted warriors upon a battle-field, and stagnation and death settle upon the spent powers of nature."[11] Similarly, the 1895 novel *The Time Machine* by H. G. Wells (1866–1946) tracks the adventures of an anonymous time-traveller who ventures 30 million years into the future, where he found the world cold, dark and nearly lifeless; now tidally locked with an expanding, cooling sun.[12] Other writers considered the future of humanity from an evolutionary perspective. For example, in First and Last Men,[13] Olaf Stapledon traces the future evolution of humanity over two billion years. He identifies eight successive species of humans during this time, the first of which is our own. The second arises from *Homo sapiens*, after the global population dwindles to 35 people who split into two groups. Although our evolutionary lineage persists, *Homo sapiens* does not.

Many of the earliest novels about human extinction focused on natural causes of disaster, although fears about science going wrong can be traced back at least to Shelley's Frankenstein.[14] The first novel to mention a technological accident destroying the world may have been Jules Verne's *Five Weeks in a Balloon*,[15] in which one character states: "I sometimes think that the end of the world will come when some immense boiler, heated to three thousand atmospheres, blows up the earth", while the first mention of a catastrophe caused by autonomous machines can be found in Samuel Butler's 1863 essay 'Darwin Amongst the Machines'.[16] By the end of World War II, the theme of scientists harnessing the sacred powers of nature to wreak unprecedented destruction had become relatively common (though they were first described in Wells' 1914 story *The World Set Free*).[17] Prominent examples of this genre include Nevil Shute's novel *On the Beach*,[18] Stanley Kubrick's film *Dr Strangelove or: How I Learned to Stop Worrying and Love the Bomb*,[19] Raymond Briggs' graphic novel *When the Wind Blows*[20] and Gudrun Pausewang's children's book *The Last Children of Schoenborn*.[21]

Writers of speculative fiction were also among the first to consider possible means of preventing global catastrophes. For instance, Lord Byron was reported to have mused with friends about the possibility of an early form of planetary defence:

> Who knows whether, when a comet shall approach this globe to destroy it, as it often has been and will be destroyed, men will not tear rocks from their foundations by means of steam, and hurl mountains, as the giants are said to have done, against the flaming mass?[22]

Similarly, William Hope Hodgson's *The Night Land* depicts humanity surviving, after the sun has burned out, in huge pyramids that are geothermally heated with crops grown underground in hydroponic rooms,[23] while the 1923 novel *Nordenholt's Million*, written by Alfred Walter Stewart under the pseudonym J. J. Cunnington, tells the story of a plutocrat who creates a refuge in Scotland after an engineered "denitrifying" bacteria causes the food supply to collapse.[24] Finally, human survival and recovery after global catastrophes is also a common literary theme. While much of this genre is not strictly concerned with existential risk, because the survival of the human species is either not in question or is not its primary focus, many works — such as E. M Forster's *The Machine Stops*,[25] Walter M. Miller Jr.'s *A Canticle for Leibowitz*,[26] Ursula K. Le Guin's *Always Coming Home*,[27] Octavia Butler's *Parable of the Sower*, Cixin Liu's *The Dark Forest*[28] and Emily St. John Mandel's *Station Eleven*[29] — remain of interest to ERS scholars.

Central themes of this body of literature include the plight of "the last man", the inevitability of some future disaster, and the folly of human hubris. According to W. Warren Wagar, science fiction was also instrumental in establishing the academic field of Futures Studies,[30] with H. G. Wells' 1901 book *Anticipations of the Reaction of Mechanical and Scientific Progress Upon Human Life and Thought* providing its foundational text,[31] followed by his Royal Institute lecture titled "The Discovery of the Future".[32] Wells argued that humanity should use the scientific method to understand how the future might unfold — in contemporary scholarly parlance, to map out the possible, probable, and preferable futures. In his words:

> And if I am right in saying that science aims at prophecy, and if the specialist in each science is in fact doing his best now to prophesy within the limits of his field, what is there to stand in the way of our building up this growing body of forecast into an ordered picture of the future that will be just as certain, just as strictly science, and perhaps just as detailed as the picture that has been built up within the last hundred years of the geological past?

Wells also wrote two non-fiction essays about the topic of human extinction, "On Extinction"[33] and "The Extinction of Man",[34] though both clearly draw as much on his literary imagination as his scientific method. Similar themes were also raised by other science fiction authors, including Arthur C. Clark, William Gibson and David Brinn. These themes are an especially noted feature of the writings of Isaac Asimov (1920–1992), a professor of biochemistry as well as a prolific popular science and science fiction author, as in his *Foundation* series concerning the predicted collapse and recovery of galactic civilisation.[35] Indeed, Asimov wrote the first book-length non-fiction treatment of possible existential catastrophes, A *Choice of Catastrophes: The Disasters That Threaten Our World* (1979).[36] Many of the science fiction authors who have had the deepest impact on ERS have frequently crossed between science fiction and science journalism or non-fiction. However, a special mention also needs to be made for the works of pure journalism that have helped to build the field. Notable examples of this include Winston Churchill's "Shall We All Commit Suicide?" in *Nash's Pall Mall Magazine*,[37] Jonathan Schell's "The Fate of the Earth" in *The New Yorker*,[38] and the anonymously written "Sui Genocide" in *The Economist*.[39]

Yet, the scientific value of this work is constrained by its commitment to storytelling and literary success. It thus focuses on apocalyptic and catastrophe narratives that readers would find engaging rather than the most plausible or realistic scenarios. Nick Bostrom has called this the "good-story bias" and warns that "if we are not careful, we can be [misled] into believing that the boring scenario is too far-fetched to be worth taking seriously".[40] Nonetheless, speculative fiction undoubtedly played a role in focusing scientific and public attention on the long-term challenges facing humanity in a hostile universe, and an early exposure to this genre of literature has also undoubtedly been a strong personal influence on many scholars in the field.

Concerned scientists

Another important contribution to the development of ERS arose from scientists who became concerned about trends and developments in their fields, which they felt might significantly harm humanity and which they wished to draw to the attention of politicians and the public.

Worries about the risk of a global catastrophe first gained major scientific attention after World War II, in response primarily to nuclear weapons. The earliest of these appears to have related to whether they might ignite the Earth's atmosphere, although these were quickly dismissed.[41] Far greater attention was given to the risk that "radioactive particles" could contaminate the environment, potentially causing a global catastrophe. This theory drew from the work of Hermann Muller, who discovered that radiation can induce genetic mutations and received the first post war Nobel Prize in physiology for this work in 1946. Muller, together with Bertrand Russell, Albert Einstein and other prominent scientists of the day, came to write in what came to be known as the *Russell-Einstein manifesto* in 1955, according to which:

> No one knows how widely such lethal radioactive particles might be diffused, but the best authorities are unanimous in saying that a war with H-bombs might possibly put an end to the human race... sudden only for a minority, but for the majority a slow torture of disease and disintegration.[42]

An important consequence of this manifesto was the establishment of the Pugwash Conferences on Science and World Affairs, which was awarded the 1995 Nobel Peace Prize for their "efforts to diminish the part played by nuclear arms in international politics and, in the longer run, to eliminate such arms". The first of these was initiated in 1957 by Russell and Joseph Rothblatt, a physicist who worked on the Manhattan Project.

Other Manhattan Project scientists established the *Bulletin of the Atomic Scientists* (*The Bulletin*) in 1945, because they were concerned about the consequences of their work. Two years later, the bulletin created the iconic "Doomsday Clock" to:

> [warn] the public about how close we are to destroying our world with dangerous technologies of our own making. It is a metaphor, a reminder of the perils we must address if we are to survive on the planet.[43]

Thus, in response to world events, *The Bulletin*'s Science and Security Board moved the minute hand toward or away from midnight, which represents global destruction. The clock was initially set to seven minutes to midnight, but in 1949 moved to five minutes to midnight and then to two minutes to midnight in 1953, after the United States and Soviet Union detonated the first thermonuclear weapons. This was the latest the clock was ever set until 2020, when the bulletin decided to move it to 100 seconds to midnight; the furthest away it has been to midnight was 17 minutes in 1991, following the end of the Cold War. Other academics had also continued working on the possibility of human extinction, such as the philosopher John Somerville, who founded the "International Philosophers for the Prevention of Nuclear Omnicide" in 1983 to "apply the resources of philosophy, in its widest sense of the term, to prevent and eliminate nuclear and other threats to global existence; create an enduring world peace; develop a just social, economic and political basis for peace and human well-being".

Worries about environmental catastrophes also emerged after the Second World War, although an awareness of humanity's profound, and potentially dangerous, impact on our environment can be traced back at least as far as the late 18th century.[44] Some of the earliest book-length studies of the potential for civilisational collapse, including William Vogt's *Road to Survival*[45] and Fairfield Osborne's *Our Plundered Planet,*[46] sounded an alarm about population growth, soil erosion and environmental pollution while also dripping with racial prejudice and colonial interests in the survival of "The West". Another pivotal early work was Rachel Carson's *Silent Spring*, which not only echoed these earlier concerns but significantly increased their scientific rigour and added a crucial policy edge by raising public awareness about the danger from chemical pesticides, such as DDT, chlordane and heptachlor.[47] Carson (1907–1964) was a marine biologist, nature writer and pioneering conservationist who became concerned about the ecological effects of indiscriminate overuse of pesticides, which she called "biocides". As she wrote in the book:

> Along with the possibility of the extinction of mankind by nuclear war, the central problem of our age has ... become the contamination of man's total environment with such substances of incredible potential for harm — substances that accumulate in the tissues of plants and animals and even penetrate the germ cells to shatter or alter the very material of heredity upon which the shape of the future depends.

In 1968, Paul (1932–) and Anne (1933–) Ehrlich, a husband and wife pair who trained as biologists but came to work predominantly in ecology and population studies, were commissioned to write *The Population Bomb*,[48] which received wide public attention. It warned about the catastrophic impacts of overpopulation, which the Ehrlichs claimed could lead to "hundreds of millions" of deaths from starvation. In 1972, the Club of Rome, an organisation of scientists, economists, diplomats, government officials, and other influencers from around the world, published a similar report called *The Limits to Growth*.[49] This developed the first global systems models to investigate the long-run impacts of trends in population, consumption, environmental degradation, and technology. Its conclusions were stark: "If the present growth trends in world population, industrialization, pollution, food production, and resource depletion continue unchanged, the limits to growth on this planet will be reached sometime within the next one hundred years".

By the early 1980s, some scientists had become worried that the greatest threat posed by nuclear conflict was not radioactivity but the massive firestorms that could inject soot into the stratosphere, blocking incoming solar radiation and causing global agricultural failures and perhaps even human extinction. The result would be what the atmospheric scientist Richard Turco called "nuclear winter". One of the most prominent scientists who warned about nuclear winter was the cosmologist, planetary physicist and exobiologist Carl Sagan (1934–96). Sagan had gained significant scientific prominence through his research, especially in the search for extraterrestrial life, and had a preeminent reputation as a science communicator through his books and TV programmes such as *Dragons in Eden*[50] and *Cosmos*.[51] Sagan and four other scientists published an influential study modelling this possibility in the journal *Science*.[52] However, Sagan also took the decision to pre-empt this publication with more popular works and media appearances to increase the potential impact of the research on politicians and the public. For instance, he wrote the cover story for the October 30th 1983 edition of *Parade*, in which he argued that, if a nuclear conflict were to occur:

> Many species of plants and animals would become extinct. Vast numbers of surviving humans would starve to death. The delicate ecological relations that bind together organisms on Earth in a fabric of mutual dependency would be torn, perhaps irreparably. There is little question that our global civilization would be destroyed. The human population would be reduced to prehistoric levels, or less. Life for any survivors would be extremely hard. And there seems to be a real possibility of the extinction of the human species.

In another article, on the policy implications of nuclear war for *Foreign Affairs*, Sagan argued that "the central point of the new findings is that the long-term consequences of a nuclear war could constitute a global climatic catastrophe".[53] Sagan and Paul Ehrlich went on to co-organise a two-day conference and co-author the 1984 book on the "long-term biological consequences of nuclear war", *The Cold and the Dark*.[54] While controversial, this scientific activism seems to have had a significant impact. For example, the Soviet Premier Mikhail Gorbachev told Ronald Reagan in 1988 that Sagan was "a major influence on ending [nuclear] proliferation".[55]

Research on the nuclear winter phenomenon was spurred in part by a study published in 1980 by Luis and Walter Alvarez.[56] This hypothesised that the non-avian dinosaurs went extinct because an asteroid struck Earth. The impact threw dust into the stratosphere, blocking out sunlight and compromising photosynthesis. The "Alvarez hypothesis", as it became known, was ground-breaking because it threatened the then-dominant paradigm that global catastrophes do not occur and the appearance of mass extinctions in the fossil record is an artefact of their incompleteness — a paradigm that had reigned since at least the 1850s. As Trevor Palmer notes, even into the late 1980s, "it was still far from clear whether mass extinctions were real events, rather than artefacts of the fossil record".[57] This changed dramatically with the (re)discovery of the Chicxulub crater on the Yucatan Peninsula in 1990, which provided sufficient evidence to convince the scientific community that global catastrophes have occurred in the past and, by implication, could occur in the future. During the 1980s, studies of volcanoes also suggested that major eruptions could also catapult particles into the stratosphere that block out incoming light. The realisation that natural catastrophes can induce mass extinctions in this way was integral to the widespread belief that anthropogenic factors, like nuclear conflict, could have similarly devastating effects.

By the early 2000s, scientists had already identified many other threats to human survival, including threats associated with artificial intelligence,[58] biological weapons,[59] nanotechnology[60] and high-energy physics experiments.[61] All these diverse threats were explored by Martin Rees (1942–) in his 2003 book, *Our Final Century: Will the Human Race Survive the Twenty-First Century*? Rees, a celebrated cosmologist who became the UK's Astronomer Royal in 1995, offered a "scientist's warning" that humanity faces unprecedented challenges in the 21st century.[62] Rees came to the gloomy conclusion that the probability of civilisation surviving the next 100 years is perhaps 50%. Although we believe that this is of little scientific or academic value, it nonetheless attracted both public and scholarly attention to existential risk issues.

Central themes of the work of concerned scientists have included the real possibility of human extinction, the risks associated with scientific and technological progress and the consequent moral responsibility of scientists for what is done with their work. Many of these scientists have also called for the creation of a form of world government, or at least for much greater government involvement in the operation of the market and the applications of scientific research. For example, in a "message to the world congress of intellectuals", Einstein declared that "mankind can only gain protection against the danger of unimaginable destruction and wanton annihilation if a supra-national organization has alone the authority to possess these weapons".[63] Others emphasised the role of scientists in informing the public about global risks. *The Bulletin* and the Pugwash Conferences exemplify this view, as does the Union of Concerned Scientists, which was founded by students and faculty at the Massachusetts Institute of Technology in 1969, to counteract the "misuse of scientific and technical knowledge presents a major threat to the existence of mankind".

However, the theoretical frameworks within which scientists work are usually relatively simplistic and tend to be useful only for linking discrete exogenous shocks with catastrophic effects; for instance, by considering a simple causal chain from nuclear conflict to firestorms to stratospheric soot to famine. We can call this the "etiological approach" to ERS. Furthermore, concerned scientists have often tended to oppose measures to reduce our collective vulnerability and exposure to the hazards they believe science might produce (such as famine relief, civil defence or geoengineering) and suggest that there is a strong trade-off, or potential for moral hazard, between such measures and reducing the risks from scientific research. This arose in part from (justifiable) worries that these measures might be ineffective, although it also seems to reflect a desire that science in general, or at least their research in particular, should only be used for beneficial rather than harmful ends. While an admirable position from which to campaign and raise awareness, this may offer an unnecessarily limited view for the purposes of risk assessment and risk management.

Section 2: Transhumanism, Utilitarianism and the Birth of ERS

While many people's work contributed to the foundation of the field of ERS, most notably John Leslie and Rees;[64] we date the beginning of Existential Risk Studies as a unified field of research to the 2002 paper "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards" by Nick Bostrom (1973–),[65] who trained in philosophy and computer science before establishing the Future of Humanity Institute within the University of Oxford's philosophy department in 2005. This work solidified a number of step-changes in thinking about existential risk and the long-term future of humanity. Whereas previous work had tended to focus on specific catastrophe scenarios or threats, Bostrom's work approached existential risk in a holistic way. Furthermore, whereas previous work focused on human extinction and civilisational collapse, Bostrom focused on catastrophes that would prevent humanity from fulfilling its potential to flourish. Human extinction is the most obvious way this could happen, but it is not the only one. For instance, if human civilisation collapsed to a state in which we could not recover culturally, economically or technologically this may be almost as bad as if we went extinct completely; even if we were to continue developing but plateau prematurely, before our peak, this could also entail a significant loss of potential for our species. Such considerations led Bostrom to define an existential risk as "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential",[66] which remains perhaps the most canonical definition of the term to date.

Maximising future value

Bostrom's novel perspective, as he presents it in the literature, is based on two normative views. The first is *utilitarianism* — in particular, a "totalist" interpretation of it. This maintains that an act is morally right if and only if it increases the total net well-being in the universe. If people have lives worth living, then the larger the population, the greater the well-being. Hence, totalist utilitarianism implies that humanity should not only strive for happiness, but create as much well-being as possible, including through the creation of as many humans with net-positive amounts of well-being as possible. This conclusion was first articulated by Henry Sidgwick, who was also the first to note that human extinction would be "the greatest of conceivable crimes from a Utilitarian point of view".[67] However, it is important to note that this principle is not universally shared, even among utilitarians; for instance, the philosopher Jan Narveson famously counters that utilitarians "are in favor of making people happy, but neutral about making happy people".[68]

But just how many humans could we create? Carl Sagan calculated that if humanity survives on earth for another 10 million years, there could come to exist some 500 trillion future people.[69] Transcending the boundaries of our planet, the Serbian astrophysicist Milan Ćirković (1971–) estimates that "the number of potentially viable human lifetimes lost per century of postponing of the onset of galactic colonization" is approximately 10^{46} — or 10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000.[70] Bostrom built on this idea in his 2003 paper *Astronomical Waste*, in which he conjectures that, if the Virgo

Supercluster contains 10^{13} stars and the habitable zone of an average star can sustain ~10^{10} biological humans, an incredible 10^{23} *biological* people per century could live in the Virgo Supercluster alone. Yet if our technologically advanced descendants opt to convert entire exoplanets into computer hardware (so-called computronium), and if this could be used to simulate human minds that would be just as valuable as our own, then some 10^{38} simulated beings with worthwhile lives could exist per century in our supercluster, Bostrom estimates. Given that there could be 10 million additional superclusters in the visible universe, it follows that the future could contain truly astronomical quantities of well-being.

Achieving humanity's potential

The second normative view, *transhumanism*, concerns a qualitative, rather than merely quantitative, element to humanity's potential future value. This is the view that humanity should not be limited by our biological nature (which transhumanists call *bio-conservatism*) but transcend it. The central tenet of transhumanism is that we should use what Mark Walker dubs "person-engineering technologies" to radically enhance our core biological features,[71] such as cognitive capacity, emotionality and healthspan, potentially resulting in the genesis of one or more species of *posthumans*.[72]

Although transhumanist themes can be found dating back to the very dawn of civilisation (they are a key theme of the *Epic of Gilgamesh*, written c. 2,000 *BCE*), it wasn't until the late 1980s and 1990s, facilitated by the internet, that a community of transhumanists formed.[73] In his 2003 paper "Transhumanist Values",[74] Bostrom writes that the "core value" of transhumanism is "having the opportunity to explore the transhuman and posthuman realms," since this could hold the key to "realiz[ing] our ideals" in ways that are presently impossible given "our current biological constitution". However, the phrase "realize our ideals" is deceptively critical as many transhumanists would see the goal of transhumanism as ushering in a techno-utopian milieu in which people become capable of realizing ideals that at present we cannot imagine. Consider Bostrom's "Letter from Utopia", in which he plays the role of a future posthuman penning a "love letter to humanity", as it were, that time-travels back to the 21st century. As the letter's author puts it, "how can I tell you about Utopia and not leave you mystified? With what words could I convey the wonder?":

> My mind is wide and deep. I have read all your libraries, in the blink of an eye. I have experienced human life in many forms and places.... You could say I am happy, that I feel good. That I feel surpassing bliss and delight. Yes, but these are words to describe human experience. They are like arrows shot at the moon. What I feel is as far beyond feelings as what I think is beyond thoughts. Oh, I wish I could show you what I have in mind! If I could but share one second with you![75]

Along similarly utopian lines, the inventor, futurist, and Google's Director of Engineering, Ray Kurzweil anticipates the exponential development of technology bringing about a history-rupturing event known as the technological "Singularity".[76] Similar views have been expressed by the AI Researcher Eliezer Yudkowsky (1979–), who founded the Machine Intelligence Research Institute (originally the Singularity Institute for AI Research) in 2005. Kurzweil and Yudkowsky were part of a conspicuously optimistic version of transhumanism called "singularitarianism" that "believes that the Singularity is possible, that the Singularity is a good thing, and that we should help make it happen".[77] In Kurzweil's words, this event is "a future period during which the pace of technological change will be so fast and far-reaching that human existence on this planet will be irreversibly altered". Driven by "the sudden explosion in machine intelligence and rapid innovation in the fields of gene research as well as nanotechnology", humanity and machine, organism and artifact, will merge into one, yielding a "world where there is no distinction between the biological and the mechanical, or between physical and virtual reality".

However, this is problematic because much of the risk facing humanity in the 21st century stems from precisely the technologies needed to achieve the goals of transhumanists and singularitarians, making these technologies "dual-use", in that they have the power to both benefit and harm. For example, CRISPR/Cas9 based techniques for gene-editing could potentially halt and even reverse ageing, but could also empower malicious agents to synthesise unnaturally dangerous pathogens. Similarly, hypothetical future devices called "nanofactories" could usher in an age of unprecedented superabundance, but could also open the door changing almost any object into any other object at very low cost. Finally, some AI experts have become increasingly concerned that a superintelligent machine could bring about the total annihilation of humanity. As Bostrom, echoing ideas from Yudkowsky, worried in 2002:

> When we create the first superintelligent entity, we might make a mistake and give it goals that lead it to annihilate humankind, assuming its enormous intellectual advantage gives it the power to do so. For example, we could mistakenly elevate a subgoal to the status of a supergoal. We tell it to solve a mathematical problem, and it complies by turning all the matter in the solar system into a giant calculating device, in the process killing the person who asked the question. [78]

As Bill Joy eloquently warned in the famous 2000 WIRED article "Why the Future Doesn't Need Us", the dangers associated with emerging technologies may be so profound that we ought "to limit development of the technologies that are too dangerous, by limiting our pursuit of certain kinds of knowledge". He goes on to suggest that instead of a "technological utopia" of some sort, we should instead aim for a society "whose foundation is altruism", in which we "conduct our lives with love and compassion for others" and where states "develop a stronger notion of universal responsibility and … interdependency".[79] Yet the only way to achieve the goals of utilitarianism and transhumanism may be to develop these very technologies. Thus, there is a need for a unified and rigorous study of how to develop these dangerous, but apparently necessary, technologies safely and beneficially. By focusing on the potential benefits of emerging technologies in the late 1990s, the potential harms gradually, and frightfully, came into focus.

The methodologies of the first wave

This, then, is the intellectual firmament out of which ERS coalesced. If one believes that the future could contain astronomical numbers of super-enhanced posthumans in a galaxy-spanning techno-utopian paradise, then one should care about every possible event that could preclude humanity from achieving that goal. As Bostrom notes, wars, epidemics, volcanic eruptions, famines, genocides and so on may ultimately be "mere ripples on the surface of the great sea of life" since "they haven't significantly affected the total amount of human suffering or happiness or determined the long-term fate of our species".[80] All that really matters are scenarios like technological stagnation, irreversible civilisational collapse and extinction. This differs significantly from previous concerns, which focused on the process of going extinct and the loss of human life, and, to account for this difference, Bostrom proposes a four-part classification of existential risks according to their outcome. These are:

1. *Bangs* — Earth-originating intelligent life goes extinct in relatively sudden disaster resulting from either an accident or a deliberate act of destruction
2. *Crunches* — The potential of humankind to develop into posthumanity is permanently thwarted although human life continues in some form
3. *Shrieks* — Some form of posthumanity is attained but it is an extremely narrow band of what is possible and desirable
4. *Whimpers* — A posthuman civilisation arises but evolves in a direction that leads gradually but irrevocably to either the complete disappearance of the things we value or to a state where those things are realised to only a minuscule degree of what could have been achieved.[81]

In every case, humanity fails to attain technological maturity in a "stable" manner, or one that would enable us to exploit our full cosmic potential. It is this novel emphasis on potentiality that leads Bostrom to formulate a heuristic to guide impersonal altruism known as the Maxipok rule, that is to: "Maximize the probability of an okay outcome, where an 'okay outcome' is any outcome that avoids existential disaster".[82]

The next question for the field of ERS to consider was how this can be achieved. Here, existential risk scholars largely fell back on the methods of their predecessors among concerned scientists (a few of whom, most notably Eric Drexler, became fully part of the ERS community). We referred to this as the "etiological approach" to understanding existential risks, its central feature being the individuation of existential risk types according to their primary causes. Example causes include supervolcanic eruptions, asteroid impacts, gamma-ray bursts, solar flares, bioengineered pandemics, ecological mass extinctions, climate change, geoengineering, self-replicating nanobots, extraterrestrial invasions and artificial general intelligence, among others. By mapping out the links from cause to catastrophe, one can devise intervention strategies to disrupt these causal chains, thereby modulating the effects. One finds this approach in both Leslie and Bostrom,[83] and it constitutes the organising principle of Bostrom and Ćirković's edited collection *Global Catastrophic Risks*, which consists of three main sections: (i) risks from nature, (ii) risks from unintended consequences, and (iii) risks from hostile acts.[84] This etiological approach offered ERS a well-defined research program for scholars to pursue: investigate the routes to disaster from triggers, and then root out the triggers to stop the disasters. Yet, this only works if there is one, or at least a relatively small number, of causal pathways that could bring about such a disaster, and if these can be modelled in a simple enough way as to allow for solutions or alternatives to be engineered. In practice, humanity has a relatively poor track record of engineering specific solutions to complex problems, although early ERS scholars like Nick Bostrom seem not to have been put off by this.[85]

Another methodological feature of this paradigm is the use of *anthropic reasoning* to obtain new information. This concerns how one should reason about one's location in space and time to gain insights into epistemically closed fields of interest, such as predicting the future and understanding other universes. One form of this reasoning is the "doomsday argument", which seeks to assess how long humanity will survive. In *The End of the World*, Leslie offers the most detailed defence to date of this argument.[86] He asks the reader to reason as if they are a random sample of all humans that will ever live. Given that there have existed between 60 and 100 billion people so far (7.8 billion of which are currently alive), the hypothesis that there will be, say, 200 billion in total is much more probable than the hypothesis that there will be 100 trillion, since it is more likely that we are near the middle of human history rather than at one extreme end or the other. Thus, the doomsday argument concludes that we are systematically underestimating the probability of human extinction in the near future. Bostrom later developed these ideas further, arguing in one case that "the doomsday argument is alive and kicking".[87] Anthropic reasoning also motivated Bostrom's "simulation argument", which purports to narrow down the space of future (and metaphysical) possibility to three scenarios: (i) humanity goes extinct relatively soon, (ii) humanity creates advanced technologies that enable us to run a large number of simulated universes but we choose not to do this, and (iii) we are almost certainly living in a computer simulation.[88] This has a number of real implications for humanity's long-term survival. For example, studies showing that we might not exist in a simulation (or that narrow down the plausible ways that we could be simulated) reduce the probability of (iii), thereby raising the probability of (i), all else being equal. While widely accepted within ERS, these arguments are generally sceptically received by outsiders.

Central themes of this paradigm thus include transcending human limitations, maximising value in the long run, building a techno-utopia, and attaining technological maturity. A primary limiting factor for this strand of research has been its commitment to transhumanism and totalist utilitarianism, which are not widely shared. If the aim of ERS is to subjugate nature, maximise economic productivity, explore the posthuman realm, and create on the order of 10^{46} future people, most people (members of the public and academics alike) are likely to conclude that the field is absurd, since they do not share these goals. While not necessarily undermining the truth of its claims, this limits both the scope of inquiry of researchers in this wave — which has focused predominantly on a small number of technology-focused risks — and the opportunities to cooperate and engage with wider communities.

Section 3: Effective Altruism, Longtermism, and the Growth of ERS

The second paradigm in ERS built on these foundations, while incorporating insights from the emerging Effective Altruism (EA) movement, which came to be embraced by the vast majority of researchers from the first wave as well as introducing many new people to the field. The EA movement is closely associated with a number of online blogs such as *Overcoming Bias* (founded in 2006 by Eliezer Yudkowsky and Robin Hanson) and *Marginal Revolution* (founded in 2003 by Tyler Cowen and Alex Tabarrok). It began to take a more substantial form after the Oxford philosopher Toby Ord co-founded *Giving What We Can*, which quickly developed chapters around the world. Ord established *Giving What We Can* after being inspired by the work of Derek Parfit, Peter Singer and others to make a personal decision to give a significant proportion of his income to charities that would most increase well-being, and receiving many enquiries from others interested in doing the same thing.

Doing the most good

The EA movement differs from the first wave of research into existential risk in having no *a-priori* commitment to transhumanism or transhumanist values. However, it is still strongly embedded within maximising the amount of value in the world (usually understood in utilitarian terms). Following Peter Singer's influential line of argument that helping someone who lives 10,000 miles away is no less ethically obligatory than helping someone drowning in a lake right in front of you,[89] the movement sees it as vitally important to find out how to do as much good as possible, regardless of whose good it is. Within EA this problem is known as "cause prioritisation", and it has traditionally been tackled via the 'NTI framework', first developed by the Open Philanthropy Project, which considers three factors:

i. How *Neglected* is the issue?

ii. How *Tractable* is the issue? and

iii. How *Important* is the issue?

Initially, the movement focused on researching and then fundraising for effective ways of alleviating global poverty (as Singer's argument suggested), most notably by fighting tropical diseases, such as malaria. However, as it developed, members raised concerns over whether this really was the most effective way to create value, and so this cause was joined by the elimination of factory farming (along with other sources of animal suffering), shaping the far future (to maximise future well-being), and most recently tackling mental illness (especially among the poor). The reason many effective altruists decided to focus on shaping the far future is that if one wants to improve the lives of as many people as possible, and if most people who will ever exist will live in the future, then one should focus on the future. This position was most extensively articulated in the philosophy PhD thesis of Nick Beckstead, who called it "longtermism".[90] As Nick Beckstead, Peter Singer and Matt Wage write:

> One very bad thing about human extinction would be that billions of people would likely die painful deaths. But in our view, this is, by far, not the worst thing about human extinction. The worst thing about human extinction is that there would be no future generations ... We believe that future generations matter just as much as our generation does. Since there could be so many generations in our future, the value of all those generations together greatly exceeds the value of the current generation.[91]

Let's break down this line of reasoning in more detail.

First, there is hardly a debate that the long-term future is neglected, both by business, governments and academics. Even more than a century after H. G. Wells first called for a serious consideration of what the future might hold, most people struggle to think about what will happen more than a few years in the future. Indeed, in the past three decades far more scholarly papers have been published about dung beetles than human extinction (Bostrom, 2013b).

Second, there are at least some reasons for thinking that improving the long-term future is tractable. The most obvious way to affect the far future of humanity is to reduce the probability of extinction, thereby ensuring that we at least have a future, and strategies to do this are readily available. Previous work using the etiological approach to risk management already discovered many potentially worthwhile risk management strategies. However, the EA-driven second paradigm expanded its focus from these "targeted" strategies, as Beckstead called them, to more indirect "broad" strategies for altering the developmental trajectory of civilisation. These include "improving education, improving parenting, improving science, improving our political system, spreading humanitarian values, or otherwise improving our collective wisdom as stewards of the future".[92]

Third, reducing the level of existential risk is clearly extremely important from the perspective of many different value systems. For example, every mainstream ethical theory seems to imply that causing (and indeed even allowing) human extinction to occur would constitute a profound moral wrong, although most do not give these wrongs the same weight that traditional utilitarianism does. Although one need not be a utilitarian to be an effective altruist, most are utilitarians or at least "most sympathetic to utilitarianism".[93] Indeed, Toby Ord has argued that utilitarianism, along with the Scientific Revolution and Enlightenment, has "greatly contributed to the upbringing of effective altruism", while the name Effective Utilitarian Community was seriously considered as an alternative name for it.[94]

Whatever the exact prevalence of utilitarianism within EA, the basic idea finds expression in the *long-term value thesis* (LTVT), which undergirds longtermism. Here the focus is broader than utilitarianism; it concerns maximising whatever one values in the world, be it art, music, poetry, science, sports, romance, and so on.[95] Since the future could be *really big*, it could contain a lot more value, and "the bigger you think the future will be, and the more likely it is to happen, the greater the value".[96] Yet, as Benjamin Todd writes, even "if you're *uncertain* whether the future will be big, then a top priority should be to *figure out* whether it will be — it would be the most important moral discovery you could make".

The NTI framework can also be used to determine which of the drivers of existential risk ERS scholars ought to focus on, implying that the biggest may not always be the best. This has led many EA longtermists to prioritise solving the "control problem" in AI safety: the problem of how to build a machine superintelligence whose value system is properly aligned with human values. This is not necessarily because EAs believe that this is the most likely way for a global catastrophe to occur, but because its combination of tractability and neglectedness (especially compared to other drivers of risk such as nuclear security and climate change) makes it an area in which the community's resources can be used most effectively. Another area in which the EA movement has tended to judge more resources were needed is global catastrophic biological risks, an area that had been paid relatively little attention by previous paradigms of ERS.

Decision theory, Bayesian reasoning and the methods of the second wave

Apart from the NTI framework, the EA community has also been greatly influenced by Expected Value Theory (EVT) and Bayesian probability, which together are seen as encapsulating the notion of applied rationality: making decisions that will maximise long-term value when one is uncertain about what to do or how things will turn out.

EVT is the most influential "decision theory" for helping agents to choose between actions that lead to uncertain outcomes. It states that rational agents should choose the action with the greatest expected value, which is calculated by averaging the probability-weighted value of every outcome that an action could produce. To quote Nick Bostrom, if 10^{54} subjective life-years could come to exist in the future, then "a mere 1% chance of [this estimate] being correct" implies that "the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth a hundred billion times as much as a billion human lives".[97] While claims such as this tend to be repeated uncritically within the EA community, their counterintuitive implications have not gone unnoticed. For instance, considerable discussion has been given to a thought experiment known as "Pascal's mugging" (after the famous Pascal's wager argument) that involves an individual who claims to be able to create immense amounts of well-being or suffering if we do, or fail to do, what they ask. Even if one were quite convinced that this individual is lying, the extremely small chance that she or he is being truthful should lead one to comply as a precaution.[98]

In 2015, Owen Cotton-Barratt and Toby Ord proposed a definition of existential risk in terms of Expected Value Theory, which differs markedly from Bostrom's canonical definition from the first wave that was based around the concept of technological maturity. They argued that Bostrom's definition failed to adequately capture catastrophes like a global totalitarian state that oppresses its citizenry for a period of time but then collapses, thus enabling humanity to continue its quest to maximise value. On their view, existential risk should refer to any "event which causes the loss of a large fraction of expected value". This definition also introduces the related concept of an "existential hope", an event that causes a large gain in expected value; the authors borrow a neologism from J. R. R. Tolkien when referring to the latter events as eucatastrophes.[99] Examples of existential hopes include designing a value-aligned machine superintelligence or becoming multi-planetary.[100]

This switch to expected value also encouraged a shift away from focusing on the avoidance of extinction events, with a growing number of EAs — most notably those affiliated with the now-defunct Foundational Research Institute (FRI) — arguing that we need to give considerable attention to the avoidance of s-risks, which are risks of outcomes that would be worse than extinction, because they contain negative values, like suffering, "on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far".[101] For example, imagine a future in which our progeny become posthuman, colonise the universe, and attain a stable state of technological maturity, thus creating vast amounts of well-being. On Bostrom's view it appears to be an OK outcome that we should work towards if we can. However, there is an important datum missing: how much suffering exists in this universe? If the answer is that the amount of suffering greatly exceeds the amount of well-being then our progeny will have realised an s-risk. Although the suffering-focused approach remains a minority position within EA, it gestures at an important insight. Many people have noted that it is difficult to adumbrate a version of utopia that anyone would actually want to live in. Yet there is probably wide agreement about what would count as dystopia: pervasive and intense pain, misery, dejection, anguish, unfulfilled desires, ignorance, loneliness, insecurity, violence, oppression, war, and genocide. One could therefore argue that we should focus on avoiding hell rather than, as the transhumanists do, reaching heaven.[102]

Bayesian probability is the view that probabilities, including those required to perform EVT, are subjective matters of belief, rather than objective facts about the universe, so agents can always assign a probability to any possible outcome, no matter how little information they have about it, and update these probability estimates to take account of new information. The rules of Bayesian probability require that no outcome, no matter how outlandish, should ever be assigned a probability of 0, because within the theory that would imply that no quantity of evidence, no matter how great, could ever persuade us to change our mind about it.

Much of the influence of Bayesian probability on ERS has been cultural, as it permits, and even encourages, the precisification of belief and the use of evidence, even very limited evidence, over more purely rationalist considerations such as the doomsday argument. However, it has also informed a number of interesting studies that seek to assess particular kinds of existential risk. These include the use of surveys that bring together expert opinions as a basis for improving risk assessments and predictions,[103] as well as toy models that assume a simple causal pathway or fixed damage distribution for different kinds of event to estimate their overall likelihood,[104] but also individual subjective judgements that simply present evidence and conclude with the author's current best guess for a given probability.[105] These methods all have a long history, but came to the fore in ERS during this second wave.[106]

As of this writing, a large portion — maybe a significant majority — of ERS scholars are effective altruists with longtermist convictions. However, perhaps the most significant contribution of EA to the field of ERS has been the influx of resources, which have significantly contributed to the movement's reputation, both academically and in popular culture. This has included providing support, both financial and intellectual, to scholars and institutions already working in the field of ERS (such as FHI and the Machine Intelligence Research Institute, MIRI), helping to found new research centres (such as the Centre for the Study of Existential Risk (2012), the Future of Life Institute (2014) and the Centre for Human Compatible AI (2016)), and supporting the work of relevant policy think tanks (such as the Nuclear Threat Initiative, the Centre of Health Security and the Centre for Security and Emerging Technologies). Of particular note has been the establishment of the Global Priorities Institute in 2018 by the Oxford Philosopher Hilary Greaves, who transitioned from researching the Philosophy of Physics to Moral Philosophy in order to increase her impact on the world. While nominally interested in all aspects of cause prioritisation, a key aspect of this centre's work has been using tools from philosophy and economics to address the epistemic, ethical and decision-theoretic challenges of trying to influence the long-term future of humanity to maximise value.

This influx of resources and talent into the field saw it expand dramatically. However, the paradigms of EA also constrain this research in several respects. For instance, many people, including most philosophers, reject the *impersonalism* that underlies effective altruism.[107] What we should care about, critics say, is not the potential well-being of currently non-existent (and possibly never-existent) possible future people, but people who exist right now. As the philosopher Amia Srinivasan writes:

> What is required [by EA] is impersonal, ruthless decision-making, heart firmly reined in by the head. This is not our everyday sense of the ethical life; such notions as responsibility, kindness, dignity, and moral sensitivity will have to be radically reimagined if they are to survive the scrutiny of the universal gaze [that utilitarianism demands]. But why think this is the right way round? Perhaps it is the universal gaze that cannot withstand our ethical scrutiny.[108]

Relatedly, instead of accepting Expected Value Theory and then concluding that existential risk reduction is very important, it is possible to reinterpret the argument that tiny reductions in existential risk are tantamount to saving huge numbers of current people as a *reduction-ad-absurdum* of the longtermist approach itself. Once again, this is not to say that the views held by most effective altruists are wrong, only that they are not so widely shared outside of the community, and this has impacted what existential risk researchers have come to see as important, neglected and tractable, as well as their ability to engage constructively with others to allocate resources to these causes. Unfortunately, it could also mean that criticisms of EA and ERS may have been self-censored out of a fear that it will lead to resources being allocated elsewhere.[109] Nonetheless, the second paradigm has clearly offered much to the development of our understanding and management of existential risks, although it remains to be seen whether longtermism has intellectual staying power.

Section 4: Systemic Complexity, Ethical Pluralism and the Diversification of ERS

Recent years have seen the emergence of a new, third wave research paradigm within ERS. Its most salient features have been its rejection of the "etiological approach" of identifying and assessing risks according to their principal direct cause, and its embrace of more substantive principles of ethical pluralism. The approach has centred on understanding the conditions and contexts within which existential risk is emerging, and on gaining a better overview of the factors that contribute to it by working with a wide range of expertise. It is thus typified by the diversity of viewpoints on issues like how to classify existential risks, what the best methods for studying them are, and how to evaluate different possible outcomes. Underlying this mosaic of opinion is a general emphasis on the complex systematicity of existential risks; that is, seeing existential risk as a phenomenon emergent from complex systems characterised by non-linear changes and feedback loops. This marks a shift away from focusing on existential hazards to considering humanity's vulnerabilities and exposure as well. This new paradigm was fostered in part by the success of the growth of ERS in attracting researchers from other fields, such as the life and earth sciences, disaster studies and public policy.

An early example of this kind of thinking can be found in a 2014 paper co-authored by Seth Baum (1980–), a risk scholar who founded the Global Catastrophic Risk Institute in 2011, and the earth system modeller Itsuki Handoh. This paper seeks to integrate the influential "planetary boundaries" framework,[110] proposed by scholars at the Stockholm Resilience Centre, with concepts from ERS. It yields a novel risk concept called the "Boundary Risk for Humanity and Nature" (BRIHN) framework that focuses specifically on the risk "of crossing a large and damaging human system threshold", where:

> crossing such a threshold could involve abrupt and/or irreversible harms to the human system, possibly sending the human system into a completely different state. The new state could involve significantly diminished populations and levels of development, or even outright extinction.[111]

Their framework is based around the twin concepts of "resilience" (humanity's ability to adapt to changes in the global systems that surround us) and its "probabilistic threshold" (the degree of change over which the risk of our resilience being insufficient to avoid an irreversible loss moves from a near impossibility to a near certainty). This important framework remains underdeveloped and only informally applied. However, it constitutes an early attempt within ERS to redirect the spotlight of scholarly attention away from epistemically neat scenarios and analyse how disasters could unfold from a perspective more grounded in "systems theory".[112] This willingness to engage with systemic complexity has helped to launch a renewed interest in catastrophic environmental risks, like climate change and loss of biosphere integrity. However, it has also had an impact on our perception of other kinds of risk. For instance, earlier waves of ERS tended to focus exclusively on the most dramatic "long-term" risks associated with the development of Artificial General Intelligence, such as the control problem. However, researchers have recently also uncovered a range of "medium-term" risks that stem from the multi-dimensional interaction between increasingly powerful AI systems and society, including concerns about the malicious use of AI.[113]

Diverging ethical approaches

Alongside efforts to more complex kinds of existential risk, this emerging group of systems thinkers have also pushed back against some canonical normative ideas within previous paradigms. For example, the assumption that developing dangerous dual-use technologies is inevitable as encapsulated by Bostrom's "technological completion conjecture", which states that "if scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained".[114] If the "default outcome" of making a value-misaligned superintelligence is "doom", then why wouldn't humanity be able to put a stop to such research (as we have been able to prevent human extinction through an act of collective suicide or a refusal to procreate)?[115] Another idea that has been scrutinized in recent years is that space colonisation constitutes an "existential panacea" that will vastly decrease the probability of extinction. For instance, Émile P. Torres (formerly Phil Torres) has argued that there are strong reasons for believing that venturing into space could have catastrophic consequences, likely causing something like an s-risk.[116] Because of these doubts, researchers of the third wave of ERS have tended to pay less attention to what the future of humanity may be like or how to ensure human "flourishing", and have instead focused more on avoiding the present risks facing humanity.

Thus, perhaps the most prominent indication that a new paradigm is forming in ERS is the growing number of researchers who are not committed to, or may even actively oppose, the notion that one should maximise future value (i.e., utilitarian ethics). One notable starting point for understanding this shift is Karin Kuhlemann's discussion of "sexy" and "unsexy" risks.[117] Kuhlemann (1979–), who is a practising lawyer and an active campaigner on population issues as well as a researcher in philosophy and public policy, observes that scholarship in ERS has so far focused almost entirely on risks with "a characteristically polarised profile: a low probability of crystallisation, perhaps very low, but should they ever crystallise, the most salient scenario — the existential outcome — has about the highest possible severity and magnitude". Such "sexy risks" exhibit three properties: first, they are *epistemically neat*, making it easy to identify which disciplines are best-suited for studying them (asteroid impacts, global pandemics, artificial intelligence, and so on). Second, they have a *sudden onset* in that they "crystallise abruptly, with obviously catastrophic outcomes from as little as a few hours to, at most, a few short years". And third, they are *technologically driven* and, as such, "have a close relationship with rather flattering ideas about human ingenuity and intellectual prowess".

Kuhlemann argues that focusing on these risks is wrongheaded. Scholars within ERS need to also consider "unsexy risks" as well. She defines these as dangerous scenarios that could produce an existential outcome, but also have a "high probability of sub-existential outcomes".[118] The three properties of unsexy risks are: first, they are *epistemically messy*, meaning that they "resist precise definition and do not ... map well onto traditional disciplinary boundaries or institutional *loci* of governance". Investigating the relevant causal factors and mitigation strategies thus requires "the combination of perspectives from multiple wildly different disciplines, which is a daunting prospect to many researchers and a poor match to how centres of research tend to be organised and funded". Second, they *build up gradually* and hence "play out in slow motion — at least as perceived by humans". This tends to "[obscure] the extent and momentum of accumulated and latent damage to collective goods, while shifting baselines tend to go unnoticed, misleadingly resetting our perception of what is normal". And finally, they are *behaviourally and attitudinally driven* in the sense that their primary causes are "the procreative and livelihood-seeking behaviours constitutive of population growth and economic growth"; these behaviours being "supported by attitudinal predispositions to oppose the kind of regulation of individual freedoms that could address the [risks] while curbing free riding". Examples include phenomena like "topsoil degradation and erosion, biodiversity loss, overfishing, freshwater scarcity, mass un- and under-employment, fiscal unsustainability, and ... overpopulation".

This emphasis on unsexy risks is motivated in part by the rejection of the futurist perspective that was an integral part of all previous paradigms of ERS and can be characterised as embracing "a techno-progressivist or transhumanism-inflected version of total utilitarianism". In contrast, Kuhlemann advocates a "normative perspective" according to which an existential catastrophe would be bad not because of the resultant opportunity cost — that is, the lost value from *Being Extinct* — but because of "the anticipated extent and severity of the harm to living, breathing human beings" that *Going Extinct* would entail.[119] When one switches from the futurist to the normative perspective, the gulf between existential and sub-existential risks collapses, which justifies a broader focus on a range of *global catastrophic risks* that include, but are not exhausted by, threats of an existential character.

The ethical paradigms of this third wave in ERS have also helped inspire more nuanced conceptions of cause prioritisation, which abandon the simplistic notion of *importance* from EA's NTI framework due to its value-ladenness in favour of a more descriptive account of what kinds of challenges most need attention. On one account, the importance of a cause is a function of three properties, namely its *significance, urgency* and *ineluctability*. The first refers to the spatiotemporal scope of the risk and who will be affected by it: the more global and transgenerational its consequences, the greater the significance. The second refers to its probable timeline of actualisation: climate change, for example, is occurring right now, whereas it seems unlikely that the technology required to create self-replicating nanobots will arrive in the next few decades (hence, climate change is more urgent). The third refers to the ostensible unavoidability of confronting the risk given the current trajectory of civilisational development. The idea is that some "risk A" that civilisation will almost certainly have to neutralise to survive should take precedence over some "risk B" that could occur but might not. Considering all three properties offers a useful methodology for quantifying a risk's importance, which renders the NTI methodology more robust. It also highlights the greater relevance of environmental and political challenges that are contemporary and unavoidable for our civilisation over potential other drivers of risk which, while neglected and tractable, are also further off, speculative and avoidable.

Risk classification and the methods of the third wave

Reflecting the lack of a single, discipline-defining, ethical perspective the third wave of ERS scholarship has tended to be less precise in its use of definitions than the previous wave. While Bostrom's canonical definitions remain popular, many now seem satisfied to refer to specific scenarios, such as "human extinction" and "civilisation collapse". Where the term *existential risk* is used it sometimes carries a rather different, more fuzzy meaning. For instance, Adrian Currie has described the term as follows:

> At base, an existential risk (X-risk) is a threat to some thing's existence.... Where many risks — catastrophic risks for instance — are understood in terms of scale (perhaps measured in terms of lives lost, or financial cost), existential risks are indexed to the set of things under that risk. Typically, the study of existential risk focuses on a narrow band of these risks, at the upper-end of the bell curve where we meet either human extinction (a species-level threat) or the loss of crucial aspects of civilization (a culture-level threat).[120]

Others have chosen to fall back on the broader concept of a *Global Catastrophic Risk* (GCR). This has been defined variously as: having "the potential to inflict serious damage to human well-being on a global scale";[121] risks that cause "significant harm" to "the entire human population or a large part thereof";[122] "possible event[s] or process[es] that, were [they] to occur, would end the lives of approximately 10% or more of the global population, or do comparable damage";[123] and "scenarios that could, in severe cases, take the lives of a significant portion of the human population, and may leave survivors at enhanced risk by undermining global resilience systems" (Avin et al., 2018). Some have even gone so far as to tailor their definitions for specific kinds of GCR; for instance, Schoch-Spana et al. define Global Catastrophic Biological Risks as "events [which] could lead to sudden, extraordinary, widespread disaster beyond the collective capability of national and international governments and the private sector to control".[124]

In order to better study these phenomena, scholars have drawn on the important early work of Baum and Handoh[125] to propose increasingly sophisticated risk assessment concepts that seek to more fully explore the space within which global catastrophes could occur and classify their salient features. The first such scheme was articulated in a 2018 paper by a highly interdisciplinary group at the University of Cambridge's Centre for the Study of Existential Risk. Shahar Avin (a philosopher of science), Bonnie Wintle (an ecologist), Julius Weitzdörfer (a disaster layer), Seán Ó hÉigeartaigh (a computational geneticist), William Sutherland (a conservation biologist) and Martin Rees (a cosmologist) begin by noting that "to date, research on global catastrophic risk scenarios has focused mainly on tracing a causal pathway from catastrophic event to global catastrophic loss of life". What is needed, then, is an exploration of "the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios that are typically addressed by single disciplines". Hence, Avin et al. propose a comprehensive framework that identifies three primary contributory factors for global catastrophes:

1) One or more *critical systems*, demarcated by "safety boundaries" that a potential threat could breach. The authors recognise seven levels of critical systems, each of which depends on the systems "below" it in a hierarchy: *sociotechnological, ecological, whole organism, anatomical, cellular, biogeochemical* and *physical*. Within each level, they identify numerous critical components, such as "stable space/time", "complex organic molecules", "viable radiation levels", and "viable temperature range" within the category of the *physical*. Similarly, the category of *sociotechnological* systems govern "climate control", "food", "health", "resource extraction", "security", "shelter" and "utilities".

2) One or more *global spread mechanisms* that enable threats to "spread globally and affect the majority of the human population". Consider the obvious but important point that the failure of a critical system, such as a regional famine, need not pose a threat to humanity if its effects are sufficiently circumscribed. As the authors write, "this separate focus on global spread allows us to identify relevant mechanisms (and means to manage or control them) as targets of study meriting further attention, and highlights interesting commonalities". Avin et al. identify three classes of spread mechanism: *natural global scale, anthropogenic networks* and *replicators*. An example of the former would be "air-based dispersal", which could enable debris from volcanic supereruptions, asteroids, comets and urban firestorms (following a nuclear conflict) to blot out the sun, thus causing worldwide crop failures. The replicators category includes not just biological entities like pathogenic viruses, but computer malware and even deleterious "memes" that hop from mind to mind across the cultural landscape.

3) Finally, one or more failures to *prevent or mitigate* either of the previous factors. This concerns our capacity to manage risk in an effective, and effectively holistic, manner. Avin et al. once again adumbrate a hierarchy of factors. First, there is the *individual* level, which includes phenomena like *cognitive biases, empowerment, motivation* and *values*. Second, there is the

> interpersonal level, which subsumes *communication, conflict resolution, connection* and *trust*. Third, there is the institutional level, which encompasses phenomena like *adaptability, decision making, ethics* and *resources*. And fourth, there is the "beyond institutional" level, which pertains to *coordination, diversity, good governance* and *representation*.[126]

Another prominent classificatory scheme has been proposed by Nick Bostrom, which relates to different kinds of "civilizational vulnerabilities" that arise from our "semi-anarchic default condition".[127] He defines this as a world order characterised by a limited capacity for preventive policing, a limited capacity for global governance, and diverse motivations among state and non-state actors. Under these conditions, Bostrom argues that our civilisation faces two classes of vulnerability (each of which can be split into two further sub-classes). However, he clearly retains the hazard-centric perspective of previous waves of ERS, and indeed labels each with an imagined technology that he feels we might be vulnerable to rather than keeping his definitions focused on the vulnerabilities themselves. The vulnerabilities he describes relate to the following scenarios:

1) Technology makes it too easy for individuals or small groups with the appropriate motivation to cause mass destruction, so that it is either:

    a) extremely easy to cause a moderate amount of harm (very easy nukes); or

    b) moderately easy to cause an extreme amount of harm (moderately easy bio-doom).

2) Technology strongly incentivises actors to use their powers to cause mass destruction, so that either:

    a) powerful actors can produce civilisation-devastating harms and face incentives to use that ability (safe first strike); or

    b) a great many actors face incentives to take some slightly damaging action such that the combined effect of those actions is civilisational devastation (worse global warming).

There is also a third class of vulnerability (referred to as type-0), which stems not from the semi-anarchic default condition of global society, but rather from our epistemic position of engaging in scientific and technological research with an imperfect understanding of what its results might be. This relates to the following scenario:

0) A technology carries a hidden risk such that the default outcome when it is discovered is inadvertent civilisational devastation (surprising strangelets).

This scheme was clearly influenced by a renewed interest in the various kinds of state and non-state actors who would either willingly (terror) or accidentally (error) destroy the world if only the means were available.[128] This concern clearly predates the modern field of ERS; however, it has been largely ignored during its formative period. For instance, Leslie considered a cluster of "risks from philosophy", as he idiosyncratically calls them, such as anti-natalism and negative utilitarianism.[129] This attentiveness to ideology was lost with Bostrom's 2002 publication, which fixated — unsurprisingly, given transhumanism's obsession with technology — almost exclusively on what we can call *technogenic* rather than *agential* threats.[130] In recent years, though, ERS scholars have once again concentrated on the agent side of the agent-artifact dyad, given that dangerous dual-use technologies (a) require *agents* or *users* to cause harm, and (b) are becoming not only more powerful but more accessible to non-state actors like small groups and even single individuals. The first such scholar to propose this was Émile P. Torres, who proposed the term "agential risk" to denote "the risk posed by any agent who could initiate an existential catastrophe in the presence of sufficiently powerful dual-use technologies either on purpose or by accident".[131] There are five basic categories of individuals/groups that give rise to agential risks, including (i) *apocalyptic terrorists*, (ii) *ecoterrorists and neoLuddites*, (iii) *omnicidal moral actors*, (iv) *idiosyncratic actors* and (v) *value-misaligned machine superintelligence*.[132] Thus, the question of "what type of individual/group would willingly push an existential-catastrophe-causing 'doomsday button' if one were within finger's reach?" has become a topic of serious scholarship only since 2017. This has further expanded the disciplinary perimeter of ERS.

Other important classificatory schemes seek to combine concepts and ideas from global catastrophic risk with those from other relevant disciplines. For instance, three scholars of disaster law and policy at the University of Copenhagen — Hin-Yan Liu, Kristian Cedervall Lauta and Matthijs Maas — combine the classification of Global Catastrophic Risk with lessons from the field of disaster studies to produce a framework for *governing boring apocalypses*.[133] This focuses on two crucial factors that have long concerned the field of disaster studies: vulnerabilities and exposures. The first refers to "propensities or weakness inherent within human social, political, economic, or legal systems, that increase the likelihood of humanity succumbing to pressures or challenges that threaten existential outcomes". The second refers to "the 'reaction surface' — the number, scope, and nature of the interface between the hazard and the vulnerability". In other words, hazards are what destroy us (a supervolcanic eruption), vulnerabilities are how we perish (global agricultural failures), and exposures are the links between the hazards and vulnerabilities (reduced incoming solar radiation around the world). However, it is not enough to merely add in these components as it can still suggest that the "existential" part of "existential risk" is associated with and only with the hazard component, which need not be the case. They thus observe that "historical studies of civilizational collapses indicate that even small exogenous shocks can destabilise a vulnerable system". It follows that there could be existential risks that are triggered by non-existential hazards but unfold as a result of "existential vulnerabilities" and/or "existential exposures".

In a similar vein, Nathan Sears has combined existential risk studies with security studies to formulate a concept of "existential security... which takes 'humankind' as its referent object against anthropogenic existential threats to human civilization and survival".[134] Finally, Cotton-Barratt, Daniel and Sandberg use public policy analysis to classify different opportunities for preventing human extinction (and other global catastrophes). These include:

1) Prevention — ensuring that events that could precipitate a global catastrophe do not occur, by identifying hazards,

understanding their dynamics, and fostering cooperation on matters of safety through dedicated institutions or beneficial customs.

2) Response — ensuring that such events do not precipitate global catastrophes, by detecting them early, reducing the time lag between detection and response, ensuring that planned responses won't be stymied by cascading impacts, and identifying leverage points to maximise their impact.
3) Resilience — ensuring that the worst effects of global catastrophes are avoided, by maintaining and increasing the diversity of human settlements and livelihoods, preparing large-scale evacuation and recovery infrastructure, and planning late-stage response measures to deploy under worst-case scenarios.[135]

Two important lessons emerge from these various frameworks. The first is that focusing only on existential hazards, while ignoring how we are vulnerable or why we are exposed to them, could actually *increase* the overall threat, because mitigating existential hazards could produce an illusion of security. As Liu, Lauta and Maas put it:

> Defeating a global pandemic, or securing mankind from nuclear war, would be historic achievements; but they would be hollow ones if we were to succumb to social strife or ecosystem collapse decades later. By proposing alternative paths that lead to existential outcomes, our taxonomy can recalibrate the calculus and reduce the prospect of an existential outcome.[136]

The second lesson is that ERS needs to expand its menu of strategies to address all the different causal factors that would be involved in bringing about an existential catastrophe. This implies that: (a) ERS should work to further diversify the academic backgrounds of researchers within the field and (b) the field should establish more effective interfaces with other disciplines that can illumine the relevant social, political, economic and technological issues.

Section 5: The Future of ERS

How might ERS evolve in coming years or decades? Here we offer a few rough-hewn thoughts.

First, the topic of existential risk will almost certainly become both less neglected by scholars and more widely known by the public, if only because of increasingly frequent environmental, biological, technological and security disasters like extreme weather, wildfires, pandemics, coastal flooding, nuclear standoffs, cyberterrorism, desertification, food supply disruptions, state shifts in the global ecosystem, economic collapse, social upheaval, political instability, cultural and religious clashes, globally orchestrated terrorism, and so on. As researchers in the field of ERS, we have seen the subject shift from being seen as crazy and outlandish to garnering mainstream attention, over just the past five years alone. As interest in the topic grows, even more media outlets will cover the day's news and, in doing so, consult with experts who may have stumbled upon the concept of existential risk and perused the corresponding literature, especially if the field successfully spreads into other disciplines. Already, *Vox Media* has a vertical, *Future Perfect*, that provides significant exposure for global catastrophic and existential risk research, while the authors of this work have also had their work reported on by (amongst others) the BBC, *The Washington Post, Vice, Quartz, The Huffington Post* and *New Scientist*. Mass movements like "Extinction Rebellion" and "Skolstrejk för Klimatet" could also make human extinction and civilisational collapse increasingly visible to the public, thereby amplifying public interest.

Second, novel existential risk scenarios could appear on the threat horizon. Consider the fact that risks associated with nuclear war, engineered pandemics, superintelligence and so on were the stuff of science fiction prior to the mid-20th century. It is likely that the majority of these new risks will relate to new technological, cultural and political developments from humanity itself. However, we certainly should not close our minds to the possibility of new kinds of natural disaster that we simply never thought about before; as Anders Sandberg, Jason Matheny, and Milan Ćirković observe, supervolcanism "was discovered only in the last 25 years, [which suggests] that other natural hazards may remain unrecognized".[137] If more existential risk scenarios are either actively created ("ontological risk multiplication") or discovered by science ("epistemic risk multiplication"), then the ranks of ERS could further swell.[138]

Third, ERS has been dominated until quite recently by a small, and relatively homogeneous, group of researchers (in terms of factors like ethnicity, gender, cultural background and social class). It is beyond question that the community is still overwhelmingly white, male, able-bodied, and English speaking and clustered around research institutes at a small number of wealthy elite universities in the USA, UK and Scandinavia (both authors of this chapter fit part of this profile), yet claims to be working for the benefit of, or even on behalf of, all humanity. This has resulted in certain issues being foregrounded more or less than they otherwise might have been if the field had been more diverse in terms of ideology, race, gender, disability and so on. For example, many marginalised peoples throughout the world do not have the luxury of engaging in armchair speculation about the astronomical value of the far future once our posthuman descendants subjugate nature, colonise the universe, and maximise economic productivity. They may even feel that it is callous for scholars steeped in the same traditions of European imperialism that already did these things to other lands and cultures, who were thus responsible for the dismal plight of so many through colonisation and slavery, to promote themselves as saviours of the human race. They may rather agree more with the sentiments of Audre Lorde's poem *A Litany for Survival* with its assertion that:

> For those of us
> who were imprinted with fear
> like a faint line in the center of our foreheads
> learning to be afraid with our mother's milk
> for by this weapon
> this illusion of some safety to be found
> the heavy-footed hoped to silence us
> For all of us
> this instant and this triumph
> We were never meant to survive[139]

Thankfully, there may be early signs that the field is diversifying, and the potential changes this diversification might bring should not be understated. With respect to gender representation, for instance, a meta-analytic reanalysis of 40 studies published in 2015 found that "men showed a stronger preference for utilitarian over deontological judgments than women when the two principles implied conflicting decisions".[140] This suggests that a more gender-diverse field might drift away from methodological habits like plugging numbers into decision-theoretic algorithms and be more interested with engaging a wider range of ethical views. Similarly, a divergence in the ethnicity and cultural background of the field may well see a return to a greater role for science fiction as an aspect of thinking about existential risk, through Afro/Asian futurisms like those of Butler (1993) and Liu (mentioned above),[141] as well indigenous futurisms, such as Daniel Wilson's (2012) *Robopocalypse* or Alexis Wright's *The Swan Book*.[142] Perhaps this diversification of thought will expose ways of thinking about existential risk that are not even conceivable to contemporary ERS scholars such as the authors of this work.

To say these things will invariably come across as criticising those who are already in this field, and of course in one sense that is what it is. However, it is not meant as a personal attack on anyone. The systems that have led to the field of ERS developing as it has are far larger than the individuals involved. Those who first imagined human extinction, like Lord Byron, Alexander Winchell, and H. G. Wells (a noted eugenicist) were deeply enmeshed within the racist hierarchy of the 19th century; the scientists who first warned about human extinction were doing so at a time when their countries were involved in political contests to determine who would dominate the world; and the scholars who were first able to unify the field of ERS into a coherent whole were, almost by necessity, those who could most easily access the financial and reputational resources of elite academic institutions. However, these arguments do strongly imply that the field not only needs to accept and embrace diversification as it naturally occurs, but that it should actively seek to diversify itself and to be a champion for a fairer and more equitable global order. While we, as ERS scholars, may have benefited hugely from the global order as it stands, it is hard to make the case that this order is in the interests of our species as a whole, and indeed it is clear that it has created institutions that are as poorly aligned with human values as any superintelligent AI that many of us fear.

Conclusion

Systematic investigation of humanity's future from a secular perspective is disappointingly novel in history. The past two decades, though, have witnessed the formation of a new field of scientific and philosophical inquiry focused on existential risks. This chapter has attempted to sketch out the historical evolution of this field from roughly 2002 until the present, with brief descriptions of the older intellectual traditions that preceded it. It argues that the field's development can be understood in terms of distinct paradigms, or waves, of research. Our aim in doing this was to add clarity to the question of why ERS took shape when it did, and how different approaches have striven to elucidate the field's central topic. At present, the two dominant, but in many ways incompatible, paradigms in this field are EA longtermism — which traces its genealogy to the futurist model — and analyses of catastrophic risk from a more systems-theoretic perspective. This chapter is written by scholars who see themselves squarely within the most recent paradigm. However, our contention is that, given the incipiency of ERS, both paradigms offer valuable insights about how we should understand, classify, and study existential risks, as well as why we should care about the topic in the first place.

Acknowledgements

We would like to thank Catherine Richardson for editorial assistance and proof-reading for this text, and Matthijs Maas, Dan Elton, David Pearce, Alexey Turchin, Azita Chellappoo, Thomas Moynihan, Luke Kemp, Apolline Taillandier and the CSER reading group for many insightful comments. This project was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation.

Notes and References

[1] Bostrom, N. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9*(1) (2002). https://www.nickbostrom.com/existential/risks.pdf. For criticism of this perspective, see Torres, E.P. 'Against Longtermism', *Aeon* (2021). https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo

[2] It has been hypothesised that the story has its origins in actual catastrophic floods such as the prehistoric flooding of the Euxin Basin (now the Black Sea) around 5,500 BCE, although this remains highly controversial.

[3] See also Torres, P. *The End: What Science and Religion Tell Us about the Apocalypse*. Pitchstone Publishing (2016).

[4] For more on these four reasons see Moynihan, T. 'Existential Risk and Human Extinction: An Intellectual History', *Futures, 116* (102495) (2020). https://doi.org/10.1016/j.futures.2019.102495

[5] Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray (1859). https://doi.org/10.1093/owc/9780199580149.003.0005

[6] Hyman, G. *A Short History of Atheism*. I.B. Tauris & Co Ltd (2010). https://doi.org/10.5040/9780755625352

[7] Byron, George Gordon Noel Byron, sixth Bar. '301 Darkness', *Lord Byron: The Complete Poetical Works, Vol. 4*, edited by Jerome J. McGann (Dec. 1816), pp. 41–460. https://doi.org/10.1093/oseo/instance.00072952

[8] Shelley, M. W. *The Last Man*. Henry Colburn (1826). https://doi.org.10.1093/owc/9780199552351.001.0001. Shelley's husband Percy also wrote the poem Ozymandius, inspired by archaeological treasures plundered from Egypt and brought to London but demanding its readers to contemplate the fact that our own 'civilisation' may one day be laid waste by the passage of time.

[9] Grainville, J. -B. F. X. C. de. *Le Dernier Homme, Ouvrage Posthume*. Deterville (1805).

[10] Winchell, A. *Sketches of Creation: A Popular View of Some of the Grand Conclusions of the Sciences in Reference to the History of Matter and of Life*. Harper & Brothers (1876). https://doi.org/10.5962/bhl.title.60805

[11] Shields, C. W. *The Final Philosophy: Or, System of Perfectible Knowledge Issuing from the Harmony of Science and Religion*. Scribner, Armstrong & Co (1877). https://doi.org/10.1037/12770-000

[12] Wells, H. G. *The Time Machine*. William Heinemann (1895). https://doi.org/10.1093/owc/9780198707516.001.0001

[13] Stapledon, O. W. *Last and First Men: A Story of the Near and Far Future*. Methuen & Co. Ltd (1930).

[14] Shelley, Mary Wollstonecraft. *Frankenstein*, edited by Nick Groom (Oct. 2019). Crossref, https://doi.org/10.1093/owc/9780198840824.001.0001

[15] Verne, J. *Cinq Semains En Ballon*. Pierre-Jules Hetzel (1863). https://doi.org/10.5479/sil.421768.39088007099849

[16] Butler, S. *Darwin Among the Machines* (June 13, 1863), revised and reprinted as 'The Book of the Machines' in Butler, S. 'The book of the machines', in Erewhon, *Or, Over*

the Range. Trübner and Ballantyne (1872). The first mention of autonomous machines causing human extinction due to value misalignment, the principal concern for many scholars of ERS, can be found in Jack Williamson's short story *With Folded Hands* (Williamson, J. 'With folded hands', in *Astounding Science Fiction*. Street & Smith (1947)) — implying that this aspect of AI ethics predates such canonical thought experiments as John Searle's *Chinese Room* (Searle, J. R. 'Minds, brains, and programs', *Behavioral and Brain Sciences 3*(3) (1980): 417–57. https://doi.org/10.1017/S0140525X00005756) or Alan Turing's *Imitation Game*, later known as the *Turing Test* (Turing, A. M. 'Computing machinery and intelligence', *Mind* LIX (236) (1950): 433–60. https://doi.org/10.1093/mind/LIX.236.433. A comprehensive record of when different kinds of AI-induced catastrophes were first described can be found at https://timelines.issarice.com/wiki/Timeline_of_AI_safety#Timeline.

17 Wells, H. G. *The World Set Free: A Story of Mankind*. Macmillan & Co. (1914). https://doi.org/10.7551/mitpress/14181.001.0001

18 Shute, N. *On the Beach*. Heinemann (1957).

19 Kubrick, S. (executive producer). *Dr Strangelove or: How I Learned to Stop Worrying and Love the Bomb*. Columbia Pictures (1964).

20 Briggs, R. *When the Wind Blows*. Hamish Hamilton (1982).

21 Pausewang, G. *Die Letzten Kinder von Schewenborn*. Otto Maier Verlag Ravensburger (1983).

22 Medwin, T. *Conversations of Lord Byron: Noted During a Residence with His Lordship at Pisa, in the Years 1821 and 1822*. Henry Colburn (1824).

23 Hodgson, W. H. *The Night Land*. Eveleigh Nash (1912).

24 Connington, J. J. *Nordenholt's Million*. Constable & Co. Ltd. (1923). https://doi.org/10.7551/mitpress/14276.001.0001

25 Forster, E. M. *The Machine Stops* (1989), cited in: *Voices from the Radium Age* (Mar. 2022), pp. 35–80. Crossref, https://doi.org/10.7551/mitpress/14183.003.0006

26 Miller, W. M. Jr. *A Canticle for Leibowitz*. J. B. Lippincott & Co. (1959).

27 le Guin, U. K. *Always Coming Home*. Harper & Row (1985).

28 Liu, C. 黑暗森林 (*The Dark Forest*). Chong Qing Chu Ban She (2008).

29 Mandel, E. S. J. *Station Eleven*. Alfred A. Knopf (2014).

30 Warren, W. W. 'H. G. Wells and the genesis of future studies', *World Future Society Bulletin*, *17*(1) (1983): 25–29.

31 Wells, H. G. *Anticipations of the Reaction of Mechanical and Scientific Progress Upon Human Life and Thought*. Chapman & Hall (1901).

32 Wells, H. G. *The Discovery of the Future* [a discourse delivered to the Royal Institution on January 24, 1902]. T. Fisher Unwin (1902).

33 Wells, H. G. 'On extinction', *Chambers's Journal* (September 30, 1893).

34 Wells, H. G. 'The extinction of man', in *Certain Personal Matters: A Collection of Material, Mainly Autobiographical*. William Heinemann (1897).

35 Asimov, I. *Foundation*. Gnome (1951); Asimov, I. *Foundation and Empire*. Gnome (1952); Asimov, I. *Second Foundation*. Gnome (1953).

36 Asimov, I. *A Choice of Catastrophes: The Disasters That Threaten Our World*. Simon & Schuster (1979).

[37] Churchill, W. S. 'Shall we commit suicide?', *Nash's Pall Mall Magazine* (September 24, 1924).

[38] Schell, J. 'The fate of the Earth', *The New Yorker* (February 1982).

[39] Anon. 'Sui genocide', *The Economist* (December 1998).

[40] Bostrom (2002).

[41] Konopinski, E. J., C. Marvin, and E. Teller. *Ignition of the Atmosphere With Nuclear Bombs*. Los Alamos National Laboratory (1946). https://fas.org/sgp/othergov/doe/lanl/docs1/00329010.pdf

[42] The full text of this manifesto can be read at https://pugwash.org/1955/07/09/statement-manifesto/

[43] Benedict, K. 'Doomsday Clockwork', *Bulletin of the Atomic Scientists* (January 2018). https://thebulletin.org/2018/01/doomsday-clockwork/

[44] Locher, F. and J. B. Fressoz. 'Modernity's frail climate: a climate history of environmental reflexivity', *Critical Inquiry, 38*(3) (2012), pp.579–98. https://doi.org/10.1086/664552

[45] Vogt, William. 'Road to survival (1948)', *The Future of Nature: Documents of Global Change*, edited by Libby Robin, Sverker Sörlin and Paul Warde. Yale University Press (2013), pp. 187–94. https://doi.org/10.12987/9780300188479-018

[46] Osborn, F. *Our Plundered Planet*. Little, Brown and Company (1948).

[47] Carson, Rachel. 'Silent spring (1962)', *The Future of Nature: Documents of Global Change*, edited by Libby Robin, Sverker Sörlin and Paul Warde. Yale University Press (2013), pp. 195–204. https://doi.org/10.12987/9780300188479-019

[48] Ehrlich, P. R, and A. H. Ehrlich. *The Population Bomb*. Ballantine Books (1968).

[49] Meadows, Donella H., Jorgen Randers and Dennis L. Meadows. 'The limits to growth (1972)', *The Future of Nature: Documents of Global Change*, edited by Libby Robin, Sverker Sörlin and Paul Warde. Yale University Press (2013), pp. 101–16. https://doi.org/10.12987/9780300188479-012

[50] Sagan, C. *The Dragons of Eden: Speculations on the Evolution of Human Intelligence*. Penguin Random House LLC (1977).

[51] Sagan, C., A. Druyan, and S. Soter (executive producers). *Cosmos: A Personal Voyage* [TV series]. PBS (1980).

[52] Turco, R. P., O. B. Toon, T. P. Ackerman, J. B. Pollack, and C. Sagan. 'Nuclear winter: Global consequences of multiple nuclear explosions', *Science* 222 (4630) (1983): 1283–92. https://doi.org/10.1126/science.222.4630.1283

[53] Sagan, C. 'Nuclear War and Climatic Catastrophe: Some Policy Implications', *Foreign Affairs* (1983). https://www.foreignaffairs.com/articles/1983-12-01/nuclear-war-and-climatic-catastrophe-some-policy-implications

[54] Ehrlich, P. R., C. Sagan, D. Kennedy, and W. O. Roberts. *The Cold and the Dark: The World After Nuclear War*. W. W. Norton & Company (1984). While Ehrlich was initially sceptical that a nuclear conflict could cause human extinction, his view eventually changed. In his words: "it was the consensus of our group that, under those conditions, we could not exclude the possibility that the scattered survivors simply would not be able to rebuild their populations, that they would, over a period of decades or even centuries, fade away. In other words, we could not exclude the possibility of a full-scale nuclear war entraining the extinction of Homo sapiens" (Badash, 2009).

55 Frances, R. M. 'When Carl Sagan warned the world about nuclear winter', *Smithsonian Magazine* (2017).

56 Alvarez, L. W., W. Alvarez, F. Asaro, and H. V. Michel. 'Extraterrestrial cause for the cretaceous-tertiary extinction', *Science* 208 (4448) (1980): 1095–1108. https://doi.org/10.1126/science.208.4448.1095

57 Palmer, T. 'Controversy catastrophism and evolution: The ongoing debate', *Springer Science & Business Media* (2012).

58 Good, I. J. 'Speculations concerning the first ultraintelligent machine', *Advances in Computers, 6* (1966): 31–88. https://doi.org/10.1016/S0065-2458(08)60418-0

59 Lederberg, J. *Biological Warfare and the Extinction of Man* [statement before the Subcommittee on National Security Policy and Scientific Developments, House Committee on Foreign Affairs] (1969).

60 Drexler, K. E. *Engines of Creation: The Coming Era of Nanotechnology*. Anchor Press/ Doubleday (1986).

61 Dar, A., A. De Rújula and U. Heinz. 'Will relativistic heavy-ion colliders destroy our planet?', *Physics Letters B*, *470*(1–4) (1999): 142–48.

62 Rees, M. *Our Final Century: Will Civilisation Survive the Twenty-First Century*? Random House (2003).

63 Einstein, A. 'A message to the World Congress of Intellectuals', *Bulletin of the Atomic Scientists, 4*(10) (1948): 295–99.

64 Leslie, J. A. *The End of the World: The Science and Ethics of Human Extinction*. Routledge (1996); Rees, 2003.

65 Bostrom (2002).

66 Bostrom (2002).

67 Sidgwick, H. *The Methods of Ethics*. Macmillan (1907). It is a mark of how recently people started seriously thinking about possible mechanisms that could bring about human extinction that Sidgwick's remark is aimed only at "[a] universal refusal to propagate the human species" that might be derived from a norm of celibacy.

68 Narveson, J. 'Moral problems of population', *The Monist* 57 (1) (1973): 62–68. https://doi.org/10.5840/monist197357134

69 Sagan (1983).

70 Ćirković, M. M. 'Cosmological forecast and its practical significance', *Journal of Evolution and Technology, 12* (2002): 1–13. http://jetpress.org/volume12/CosmologicalForecast.htm.

71 Walker, M. 'Ship of fools: Why transhumanism is the best bet to prevent the extinction of civilization', *The Global Spiral* (2009).

72 Bostrom, N. 'Why I want to be a posthuman when I grow up', *Medical Enhancement and Posthumanity*, 107–36. Springer (2008b).

73 Although the favoured term was initially "extropianism" — where *extropy* is meant to contrast with *entropy* (see More, M. *Principles of Extropy*. Extropy Institute (2003); Bostrom, N. 'A history of transhumanist thought', *Journal of Evolution and Technology, 14*(1) (2005). https://www.nickbostrom.com/papers/history.pdf. Incidentally, Max More, who was a prominent extropian, was born "Max O'Connor", but changed his name. As he explains: "It seemed to really encapsulate the essence of what my goal is: always to improve, never to be static. I was going to get better at everything,

become smarter, fitter, and healthier. It would be a constant reminder to keep moving forward" (Regis, E. 'Meet the extropians', *Wired* (October 1994). https://www.wired.com/1994/10/extropians/

74 Bostrom, N. 'Transhumanist values', *Ethical Issues for the 21st Century*, edited by F. Adams. Philosophical Documentation Center Press (2003b).

75 Bostrom, N. 'Letter from Utopia', *Studies in Ethics, Law, and Technology, 2*(1): 1–7 (2008a). https://doi.org/10.2202/1941-6008.1025

76 Kurzweil, R. *The Singularity Is Near: When Humans Transcend Biology*. Viking (2005).

77 Yudkowsky, E. 'The Singularitarian Principles v.1.0.2', Yudkowsky.net (2001). Currently accessible via https://web.archive.org/web/20081229202843/ http://www.yudkowsky.net/obsolete/principles.html. Note that Yudkowsky disavows "everything [I wrote from] 2002 or earlier" and that these principles are no longer available from their orrigional source. However, they can still be accessed via the Wayback Machine at https://web.archive.org/web/20081229202843/ http://www.yudkowsky.net/obsolete/principles.html

78 Bostrom (2002). Yudkowsky refers to this risk as "subgoal stomp"; see Yudkowsky, E. *Creating Friendly AI 1.0: The Analysis and Design of Benevolent Goal Architectures*. The Singularity Institute (2001).

79 Joy, B. 'Why the future doesn't need us', *Wired* (April 2000). These recommendations may be seen as prescient of how the field of ERS would develop in its "second wave", although there is no obvious connection between the two.

80 Bostrom (2002).

81 Bostrom (2002).

82 Bostrom (2002). This is not to say that there weren't antecedents in the literature that deserve credit. For example, Derek Parfit famously argued that the difference between 99 percent and 100 percent of humanity dying out is far greater than the difference between 1 percent and 99 percent dying out because the former would entail a permanent end to the human story but the latter might not, e.g., if civilisation manages to rebuild (Parfit, D. *Reasons and Person*. Oxford University Press (1984)). And the calculation above from Sagan that 500 trillion future people could exist was propounded in a 1983 article about nuclear winter, which emphasised that "if we are required to calibrate extinction in numerical terms, I would be sure to include the number of people in future generations who would not be born." Yet both Parfit and Sagan focused on extinction in particular, while the original paradigm within ERS recognised that wholly survivable scenarios — even scenarios in which we attain technological maturity — could *still* result in "existentially catastrophic" outcomes.

83 Leslie (1996); Bostrom (2002).

84 Bostrom, N. and M.M. Ćirković. *Global Catastrophic Risks*. Oxford University Press (2008).

85 Sandberg, A. and N. Bostrom. *Global Catastrophic Risks Survey*. Future of Humanity Institute, University of Oxford (2008). https://www.fhi.ox.ac.uk/reports/2008-1.pdf

86 Leslie (1996).

87 Bostrom, N. 'The Doomsday Argument Is Alive and Kicking', *Mind, 108*(431) (1999): 539–51. https://doi.org/10.1093/mind/108.431.539

88 Bostrom, N. 'Are we living in a computer simulation?', *The Philosophical Quarterly, 53*(211) (2003a): 243–55. https://doi.org/10.1111/1467-9213.00309

89 Singer, Peter. 'Famine, affluence, and morality', *Philosophy & Public Affairs* (1972): 229–43.

90 Beckstead, N. *On the Overwhelming Importance of Shaping the Far Future* [PhD thesis]. Department of Philosophy, Rutgers University (2013).

91 Beckstead, N., P. Singer, and M. Wage. 'Preventing human extinction', *Effective Altruism Forum* (August 2013). https://forum.effectivealtruism.org/posts/tXoE6wrEQv7GoDivb/preventing-human-extinction

92 Beckstead (2013).

93 An increasingly common concern among effective altruists is the problem of "normative uncertainty", that humanity is currently not in a position to make absolute statements about ethical value. This has led many in the movement to eschew any statements of absolute commitment to an ethical view, preferring to state their degree of personal credence in it (i.e., their current assessment of the likelihood of its truth). See MacAskill, W. *Normative Uncertainty* [PhD thesis]. University of Oxford (2014).

94 Ord, T. and W. MacAskill. 'Opening keynote', *Effective Altruism Global 2016*. Centre for Effective Altruism. https://www.youtube.com/watch?v=VH2LhSod1M4

95 Eliezer Yudkowsky referred to this idea of promoting whatever is valuable as 'Fun Theory'.

96 Todd, B. 'Introducing longtermism', *80,000 Hours* (October 2017). https://80000hours.org/articles/future-generations/

97 Bostrom, N. 'Existential risk FAQs', *Existential Risk: Threats to Humanity's Future FAQs* (2013a). https://www.existential-risk.org/faq.pdf

98 Yudkowsky, E. 'Pascal's mugging: Tiny probabilities of vast utilities', *LessWrong* (2007a). https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities

99 Tolkien, J.R.R. *On Fairy-Stories*. Oxford University Press (1947).

100 Cotton-Barratt, O. and T. Ord. *Existential Risk and Existential Hope: Definitions*. Future of Humanity Institute, University of Oxford (2015). https://www.fhi.ox.ac.uk/reports/2015-1.pdf. For more on the different definitions of existential risk see Torres, P. 'Existential risks: A philosophical analysis', *Inquiry* (2019), pp.1–26.

101 Althaus, D. and L. Gloor. *Reducing Risks of Astronomical Suffering: A Neglected Priority*. Center on Long-Term Risk (September 2016). https://longtermrisk.org/reducing-risks-of-astronomical-suffering-a-neglected-priority/

102 See also Popper, K. *The Open Society and Its Enemies*. Routledge (1945). A potentially more useful concept is that of *protopia*, which Kevin Kelly (2011) defines as "a state that is better than today than yesterday, although it might be only a little better."

103 Müller, V. C. and N. Bostrom. 'Future progress in artificial intelligence: A survey of expert opinion', in *Fundamental Issues of Artificial Intelligence*, edited by V. C. Müller. Springer (2014). https://www.nickbostrom.com/papers/survey.pdf; Sandberg and Bostrom (2008).

104 Millett, P. and A. Snyder-Beattie. 'Existential risk and cost-effective biosecurity', *Health Security, 15*(4) (2017): 373–83. https://doi.org/10.1089/hs.2017.0028; Snyder-Beattie, A. E., T. Ord and M. B. Bonsall. 'An upper bound for the background rate of human extinction', *Scientific Reports, 9*(1) (2019): 11054. https://doi.org/10.1038/s41598-019-47540-7

105 Pamlin, D. and S. Armstrong. *Global Challenges — Twelve Risks That Threaten Human Civilisation — The Case for a New Category of Risks*. Global Challenges Foundation (2015); Ord, T. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing (2020).

106 See also Tonn, B. and D. Stiefel 'Evaluating methods for estimating existential risks', *Risk Analysis, 33*(10) (2013): 1772–87. https://doi.org/10.1111/risa.12039 and Beard, S., T. Rowe and J. Fox. 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards', *Futures, 115* (2020): 102469. https://doi.org/10.1016/j.futures.2019.102469

107 Bourget, D. and D. J. Chalmers. 'What do philosophers believe?', *Philosophical Studies, 170* (2014): 465–500. https://doi.org/10.1007/s11098-013-0259-7

108 Srinivasan, A. 'Stop the robot apocalypse', *London Review of Books* (September 2015). https://www.lrb.co.uk/the-paper/v37/n18/amia-srinivasan/stop-the-robot-apocalypse

109 Naturally, such claims are hard to verify. However, one possible case was put together by Simon Knutsson; see https://www.simonknutsson.com/problems-in-effective-altruism-and-existential-risk-and-what-to-do-about-them/. We refer to this case without commenting on the veracity of the allegations Knutsson raises.

110 Rockström, J., W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. F. Lambin, T.M. Lenton et al. 'A safe operating space for humanity', *Nature, 461* (7263) (2009): 472–75. https://doi.org/10.1038/461472a

111 Baum, S. D. and I. C. Handoh. 'Integrating the planetary boundaries and global catastrophic risk paradigms', *Ecological Economics, 107* (2014): 13–21. https://doi.org/10.1016/j.ecolecon.2014.07.024

112 Another influential early article was co-authored by Seth Baum and a group of scholars who attended a month-long conference on global catastrophic risk at the University of Gothenburg. This explored four possible future trajectories of civilisation, namely, "(1) Status quo trajectories, in which human civilization persists in a state broadly similar to its current state into the distant future; (2) Catastrophe trajectories, in which one or more events cause significant harm to human civilization; (3) Technological transformation trajectories, in which radical technological breakthroughs put human civilization on a fundamentally different course; (4) Astronomical trajectories, in which human civilization expands beyond its home planet and into the accessible portions of the cosmos". Baum, S. D., S. Armstrong, T. Ekenstedt, O. Häggström, R. Hanson, K. Kuhlemann, M. M. Maas et al. 'Long-term trajectories of human civilization', *Foresight, 21*(1) (2019): 53–83. https://doi.org/10.1108/FS-04-2018-0037

113 Prunkl, C. and J. Whittlestone. 'Beyond near- and long-term: Towards a clearer account of research priorities in AI ethics and society', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (*AIES'20*) (2020), pp.138–43. https://doi.org/10.1145/3375627.3375803

114 Bostrom, N. 'The future of humanity', in *New Waves in Philosophy of Technology*, edited by J. -K. B. Olsen, E. Selinger, and S. Riis. Palgrave McMillan (2009). https://www.nickbostrom.com/papers/future.pdf. See also Kurzweil (2005).

115 Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (2014).

116 Torres, P. 'Space colonization and suffering risks: Reassessing the "Maxipok Rule"', *Futures, 100* (2018): 74–85. https://doi.org/10.1016/j.futures.2018.04.008. For a

contrary view, see Ćirković, M. M. 'Space colonization remains the only long-term option for humanity: A reply to Torres', *Futures, 105* (2019): 166–73. https://doi.org/10.1016/j.futures.2018.09.006

117 Kuhlemann, K. 'Complexity, creeping normalcy and conceit: Sexy and unsexy catastrophic risks', *Foresight, 21*(1) (2019): 35–52. https://doi.org/10.1108/FS-05-2018-0047

118 Kuhlemann (2019).

119 See Chapter 7 of Torres, P. *Human Extinction: A History of the Science and Ethics of Annihilation*. Routledge (2023).

120 Currie, A. 'Existential risk, creativity & well-adapted science', *Studies in History and Philosophy of Science Part A, 76*: 39–48 (2019). https://doi.org/10.1016/j.shpsa.2018.09.008

121 Bostrom and Ćirković (2008).

122 Bostrom, N. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (2013): 15–31. https://doi.org/10.1111/1758-5899.12002

123 Cotton-Barratt, O., S. Farquhar, J. Halstead, S. Schubert and A. Snyder-Beattie. *Global Catastrophic Risks* (2016).

124 Schoch-Spana, M., A. Cicero, A. Adalja, G. Gronvall, T. Kirk Sell, Di. Meyer, J. B. Nuzzo et al. 'Global catastrophic biological risks: Toward a working definition', *Health Security, 15*(4) (2017): 323–28. https://doi.org/10.1089/hs.2017.0038

125 Baum and Handoh (2014).

126 Avin, S., B. C. Wintle, J. Weitzerdörfer, S. S. Ó hÉigeartaigh, W. J. Sutherland and M. J. Rees. 'Classifying global catastrophic risks', *Futures, 102* (2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001

127 Bostrom, N. 'The vulnerable world hypothesis', *Global Policy, 10*(4) (2019): 455–76. https://doi.org/10.1111/1758-5899.12718

128 So far as we know, this terror/error distinction originated from Rees (2003).

129 Leslie (1996).

130 Nonetheless, Martin Rees pays some attention to the threats posed by a "lone dissident or terrorist" and "embittered loners and dissident groups" (Rees, 2003; see also Torres, P. *Morality, Foresight and Human Flourishing: An Introduction to Existential Risks*. Pitchstone Publishing (2017)).

131 Torres, P. 'Facing disaster: The great challenges framework', *Foresight, 21* (1) (2019): 4–34.

132 Torres (2018); Torres (2019). See Torres, P. 'Maniacs, misanthropes, and omnicidal terrorists: Reassessing the agential risk framework', *Proceedings of the Stanford Existential Risks Conference* (forthcoming).

133 Liu, H. Y., K. C. Lauta and M. M. Maas. 'Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research', *Futures, 102* (2018): 6–19. https://doi.org/10.1016/j.futures.2018.04.009

134 Sears, Nathan Alexander. 'Existential security: Towards a security framework for the survival of humanity', *Global Policy, 11*(2) (2020): 255–66. https://doi.org/10.1111/1758-5899.12800

135 Cotton-Barratt, O., M. Daniel and A. Sandberg. 'Defence in depth against human

extinction: Prevention, response, resilience, and why they all matter', *Global Policy, 11*(3) (2020): 271–82. https://doi.org/10.1111/1758-5899.12786

136 Liu, Lauta and Maas (2018).

137 Sandberg, A., J. G. Matheny and M. M. Ćirković. 'How can we reduce the risk of human extinction?', *Bulletin of the Atomic Scientists* (September 2008). https://thebulletin.org/2008/09/how-can-we-reduce-the-risk-of-human-extinction/

138 Torres (2016); Torres (2017).

139 Lorde, A. *A Litany for Survival* (pp. 31–32). Blackwells Press (1981).

140 Friesdorf, R., P. Conway and B. Gawronski. 'Gender differences in responses to moral dilemmas: A process dissociation analysis', *Personality and Social Psychology Bulletin, 41*(5) (2015): 696–713. https://doi.org/10.1177/0146167215575731

141 Butler, O. E. *Parable of the Sower*. Four Walls Eight Windows (1993); Liu (2008).

142 Wilson, D. H. *Robopocalypse*. Knopf Doubleday Publishing Group (2011); Wright, A. *The Swan Book*. Giramondo Publishing (2013); Mitchell, A. and A. Chaudhury. 'Worlding beyond "the 'end' of 'the world'": White apocalyptic visions and BIPOC futurisms', *International Relations, 34*(3) (2020), pp.309–32.

2. Democratising Risk: In Search of a Methodology to Study Existential Risk

Carla Zoe Cremer and Luke Kemp

Highlights:

- Existential Risk Studies is currently dominated by a Techno-Utopian Approach (TUA) that defines existential risk, not in terms of human extinction, but in terms of the loss of very large quantities of value predicated on a utilitarian, transhumanist, and longtermist set of ethical assumptions.
- The TUA is not representative of the values of most human beings. This is philosophically tenuous, given that the field often speaks of humanity as a whole, and also languorously undemocratic, given its alignment with the interests of elites who are also contributing to existential risk.
- Defining existential risk in relation to the loss of value invariably ties the field to a particular group's value system. To avoid this, we should separate the study of existential and global catastrophic risk from both extinction ethics (what is good or bad about human extinction) and existential ethics (what is good or bad about different societal forms).
- The TUA also presents methodological limitations including an apparent commitment to technological determinism, simplistic, threat-based models of risk assessment, and expected value-based approaches to decision-making under

https://doi.org/10.11647/OBP.0360.02

uncertainty. These are hard to justify and push the field away from best practise in related fields such as disaster risk reduction and climate science.

- Given the above, Existential Risk Studies currently poses a high level of "response risk", i.e. it could recommend responses that are net harmful to humanity, with many historical precedents for how this can happen. To prevent this existential risk, scholars should transparently acknowledge the moral and empirical assumptions used in risk analyses; critically embrace the latest advances in risk assessment from other fields; diversify the field's approaches and assumptions; and democratise its judgement.

The importance of democracy in the governance of Existential Risk research is also discussed in Chapter 16, while mechanisms for engaging with the democratic process in policy making are discussed in Chapter 22.

1. Introduction

Over the past two decades, scholars have begun to methodically study human extinction and global catastrophes. This field of "Existential Risk Studies" (ERS) aims to (i) identify existential and catastrophic risks; (ii) map out the potential causes of existential catastrophes; (iii) understand the ethical implications of such calamities and (iv) devise effective strategies for mitigation and prevention.

Although the field is relatively small, it has expanded considerably, especially over the past 10 years.[1] It is also of increasing public interest: several popular trade-books have been published.[2] The ideas have been integrated into vision-setting reports from the UN Secretary General.[3] Institutions focusing on existential risk have received hundreds of millions of dollars in philanthropic funding.[4] It is a field on the rise.

It is commendable and overdue that the study of human extinction is receiving greater academic engagement and public attention. However, this field needs to be held to high standards: it is ambitious, could affect the lives of many and attracts scholars who seek to change the trajectory of global society.

The field faces daunting challenges. How can it be inclusive of the diversity of human preferences and visions of the future? How can researchers avoid baking their subjective assumptions into risk analyses that might affect those who do not share their values? How do they conduct complex risk assessments? How do they deal with uncertainty? How do they compare risks with different quantities of evidence and degrees of plausibility? How do they ensure that the catastrophes the field studies are not misused to justify dangerous actions?

The field has not yet established the answers to such questions and we are not the first to be aware of this.[5] Throughout this chapter we will point to the scholars who we know have raised similar questions. The historically dominant Techno-Utopian Approach (henceforth the "TUA") played an important role in establishing the field and drawing attention to the significance of studying human extinction. It is time to examine this approach with a critical eye. We do this to identify weaknesses, areas for further investigation and the need to also explore alternative approaches. The TUA, which relies heavily on total utilitarianism, transhumanism, and (strong) longtermism, is too morally unrepresentative, methodologically flawed, and risky an approach to rely on. It is time to diversify the definitions, tools, and frameworks of ERS. We need to develop an ERS methodology that addresses the core questions of ERS and avoids the problems of the TUA.

The question we raise is: what should the study of human futures and catastrophe look like under moral uncertainty? We suggest some solutions: a diverse range of approaches, deliberative democratic processes, and the separation of the study of catastrophe and extinction from the ethics of human existence and extinction.

We proceed in Section 2 by outlining the moral assumptions of the TUA. In Section 3 we give reasons why the TUA is not representative of wider human preferences and why representation matters to the goals of ERS. Section 4 explains how moral and empirical assumptions are often masked by abstract and ambiguous definitions and tools. Moral and empirical assumptions are important because they can distort the results of our work, whether it be in how we conceive of risks, which risks we prioritise, or what policies we recommend. In Section 5 we focus on how studying existential risk could backfire by growing other sources of catastrophic risk. In Section 6 we suggest a democratic, risk-averse approach.

2. The Techno-Utopian Approach to ERS

We focus on the Techno-Utopian Approach to existential risk for three reasons. First, it serves as an example of how moral values are embedded in the analysis of risks. Second, a critical perspective towards the Techno-Utopian Approach allows us to trace how this meshing of moral values and scientific analysis in ERS can lead to conclusions, which, from a different perspective, look like they in fact increase catastrophic risk. Third, it is the original and by far most influential approach within the field.

2.1 Definitions and history

2.1.1 The influence of the TUA

The TUA is a cluster of ideas which make up the original paradigm within which the field of ERS was founded. We understand it to be primarily based on three main pillars of belief: transhumanism, total utilitarianism and strong longtermism. More precisely: (1) the belief that a maximally technologically developed future could contain (and is defined in terms of) enormous quantities of utilitarian intrinsic value,[6] particularly due to more fulfilling posthuman modes of living;[7] (2) the failure to fully realise or have capacity to realise this potential value would constitute an existential catastrophe;[8] and, (3) we have an overwhelming moral obligation to ensure that such value is realised by avoiding an existential catastrophe,[9] including through exceptional actions.[10]

Not all publications that make use of the TUA explicitly support every element of the approach, but the most widely read publications incorporate a significant number of TUA elements and share its visions of the long-term future.[11] The most popular definitions of existential risk are still the initial techno-utopian definitions by Bostrom and more abstract, but very similar versions based on expected value (see Section 4.1). The few attempts to put forward alternative frameworks[12] are not nearly as widely cited, known, or used.[13] Importantly, they do not offer alternative definitions of what an existential risk is. Despite the increase in size and diversity in the field of ERS there appears to still be no coherent alternative to the TUA.

The impact of the ideas of the TUA can be seen across the field, characterising its most cited and best-known publications. Beard and Torres trace the beginning of the existing field of ERS to the early publications by Bostrom.[14] It was Bostrom's work in the early 2000s that first aimed to formalise the concept of an existential risk, notably via his 2002 paper *"Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards"*[15] and the 2013 article *"Existential Risk Prevention as Global Priority"*,[16] which have been cited a combined 1,046 times.[17] These two papers articulate the canonical definitions of "existential risk" and, along with Bostrom's 2003 paper *"Astronomical Waste"*[18] present the clearest distillation of the TUA.

The TUA also characterises almost every existential risk text with significant public profile. This includes trade-books such as *Superintelligence,*[19] *The Precipice,*[20] *Life 3.0,*[21] and *What We Owe the Future.*[22] Several culturally influential ideas such as a technological singularity[23] and longtermism[24] are influenced by the TUA. The techno-utopian worldview also appears to resonate with key funders. For example, Holden Karnofsky, the co-founder and co-chief executive of Open Philanthropy, strongly echoes the TUA in his "Most Important Century" series.[25]

The TUA we describe should be understood as an ideal type:[26] both the texts and thinkers under it may vary in specifics but converge in their broad vision. This is not an issue: despite many national variations we can speak of capitalism in general and recognise particular countries as being capitalist.[27]

2.1.2 Defining existential risk under the TUA

Bostrom provides two general formulations of existential risk. He initially defined it as "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential".[28] Later, he provided a more refined definition: "one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development".[29] The latter paper provides a further typology of existential risks as:

i. Premature extinction of humankind before reaching technological maturity;
ii. Failure to reach technological maturity due to an unrecovered collapse, recurrent collapse, or plateauing;
iii. Technological maturity being realised in an irredeemably flawed manner; or,
iv. Ruination of Earth-originating intelligent life after technological maturity is reached but before its full potential can be realised.

According to this typology, the core feature of an existential catastrophe is a failure to attain a stable state of technological maturity, maintained for as long as is physically possible. Bostrom specifies this as a level of technological development, resource acquisition, and resource efficiency that allows for the highest feasible level of "economic productivity and control over nature".[30] Failure is anything but the full exploration of possible options and the full exploitation of available matter.

A recent definitional reworking of this approach is the formulation of an existential catastrophe as "an event which causes the loss of a large fraction of expected value"[31] or less technically as a risk that "threatens the destruction of humanity's longterm potential".[32]

2.2 Transhumanism: Humans as a stepping stone

Transhumanism is the moral position that there is value in exploring posthuman and transhuman modes of being.[33] The results are to be beings — modified biological humans, cyborgs, androids, or digital simulants — whose lives are considered more valuable than current ones.[34] Transhumanists argue that these beings could achieve far longer, richer lives marked by net positive experiences.[35]

Achieving such lives would depend on further technological progress.[36] We would need three fundamental transformations: (i) protecting life; (ii) expanding cognition, and (iii) elevating well-being.[37] This can, for example, take the form of achieving immortality, superintelligence, and a greater capacity for pleasure.

Although transitioning to a posthuman stage could of course entail the extinction of *Homo sapiens*, Bostrom contends that "the permanent foreclosure of any possibility of this kind of transformative change of human biological nature may itself constitute an existential catastrophe".[38] Preventing existential risk is not primarily about preventing the suffering and termination of existing humans; it is focused on preserving humans so that they may give rise to a posthuman species that contains more value.

2.3 Total utilitarianism: Humans as containers

Total utilitarianism identifies moral rightness with the maximisation of well-being. Well-being could be interpreted in hedonistic, desire-satisfactionist, or objective-list theory terms.[39] People thus carry some unit of value, and the greater this value, the better. Total utilitarianism therefore demands that we maximise the total amount of value in the universe, with as many people coming to exist as possible, each person living an overall happy (i.e., net-positive) life, regardless of where or when these people come into existence. This equivalence in moral patient-hood between different "containers" of value here relies on the "impersonalist" or "non-identity" perspective, in which it is not relevant who is affected, only that someone is affected.

The utilitarian argument that the future should be an overwhelming moral priority relies on an assumption that the number of intelligent beings who could come to exist could be unimaginably large. Matheney, for example, estimates a low-range figure of 10^{16}, assuming we remain Earth-bound.[40] Bostrom argues that if our descendants colonised as much of the universe as quickly as possible and converted celestial bodies into "computronium" for running simulations, this could result in some 10^{38} simulated conscious beings per century in Earth's supercluster alone.[41] Assuming wider interstellar exploration then we could even produce 10^{58} happy simulations.[42] If we are alone in the universe and computations are run at colder temperatures towards the end of the heat death of the universe then even more could be achieved.[43]

The posthuman calculus, or the "astronomical value" thesis, is that the future could hold far greater value than the present due to the teeming masses of beings with higher levels of happiness than organic mortals can achieve. *Homo sapiens* is eclipsed by the towering shadow of the techno-utopian future.

2.4 Strong longtermism: The present in the shadow of the future

The third philosophical foundation is strong longtermism. Considering future generations as moral patients is by no means new. Notions of intergenerational justice, equity and fairness across the deep future have been extensively discussed for decades.[44] Strong longtermism goes a step beyond this and suggests that for some situations we may have an ethical imperative to select the choices expected to have the best effect on the long-run future,[45] and usually relies on a utilitarian calculus to justify this.

The best choice is often equated with the choice that has the highest expected value.[46] Expected value is calculated by multiplying the value of an outcome by the probability of it occurring. A calculus that numerically favours strong longtermist actions, such as reducing existential risk, rather than saving millions of today's people, often relies on the assumption of continued technological development, happy future people, and interstellar settlements.[47] Thus, ensuring technological progress and maximising the quantity and expected quality of hypothetical future lives may be deemed more important than protecting current lives. We have no principled guidance about when and why a strong longermist should prioritise living humans of today.

Transhumanism, total utilitarianism, and strong longtermism are a coherent, re-enforcing and complementary set of beliefs. Some versions of longtermism might be compatible with multiple theories of value,[48] but this is an ongoing area of study.[49] It is unclear how different those versions would be and what this would practically imply.[50] The reasons put forward for why other moral theories should care about existential risk include to cultivate civilisational virtues and/or to meet intergenerational obligations.[51] These reasons explain why different moral theories should be concerned about human extinction, but not why they would support strong longtermism.

The non-utilitarian case for strong longtermism is, for now, weak. There have been too few attempts to understand whether different moral positions can support the belief that the vast majority of value lies in the future. There is countering evidence that this is not the case.[52] While it may be the case that a wide range of moral positions support caring about the future in a broad sense, it is not necessarily the case that each of these views yield the same practical implications. Hence, it is not appropriate to argue that the implications of strong longtermism follow from a diverse range of moral positions.

3. Representation and Existential Risk

3.1 The TUA as a non-representative view

The TUA is not representative of what most humans alive now believe. Relying on the TUA, which is unrepresentative of many people's moral views today, can distort the analysis of existential risk. Representativeness itself says nothing about whether its philosophical pillars are wrong or right, but it is risky to rely exclusively on one unrepresentative approach given moral, political and empirical uncertainty. Theoretical work in ERS should be paired with and constrained by empirical studies that capture the range of existing intuitions about human extinction and longtermism. We must know the moral intuitions of the public, and when experts dismiss their moral intuitions as incorrect, they must have strong arguments to do so.

There is no consensus among philosophers on moral theory. Utilitarianism is not the most commonly held view In one of the few surveys in the area less than a quarter (23.6%) of philosophers identified with consequentialism.[53] An even smaller number will be utilitarian, and a small number still will be total utilitarian. Techno-utopia offers futures of pleasurable, often virtual experiences, in which commonly valued attributes like purpose, virtue, love, and justice do not play a central role.

Transhumanism, too, is a niche perspective, and surveys reveal that those who identify as transhumanists come from a narrow demographic. The most recent high-quality survey, one that collected 760 responses from members of the World Transhumanist Association (now called Humanity+) in 2007, found that 90% of the respondents were male with a median age of 30–33 years old.[54] It is unknown how many people of the wider population would accept all or some of the premises of transhumanism if they were surveyed.

The implications of the TUA definition of existential risk also appear to be unrepresentative. According to the original definition of existential risk (a failure to attain technological maturity) technological plateauing — or a failure to spread beyond Earth — is an existential risk. Both near-term extinction due to nuclear war and a future in which humans persist sustainability and equitably for the next billion years without major technological progress are seen as existential risks: worst-case outcomes for humanity. Equating these outcomes as morally equivalent is likely unintuitive to many.[55]

The perspective that potential future lives are morally equivalent to existing lives may also be unintuitive to many. It is an active area of theoretical debate between philosophers, and we need more surveys that empirically query the moral intuitions of a wider population. Caviola et al.[56] find context-dependent support for adding future people. Schubert et al.[57] find context-specific overlap between surveyed intuitions of lay-persons and theoretical arguments by experts in ERS, although this depends heavily on survey framing. These are a commendable start. However, far more work is needed to understand what the wider population of the world wants from the far future. Ideally, this should be built not just on surveys, but also more deliberative practices (see Section 6).

The techno-utopian vision of the future, which combines three rather uncommon positions (transhumanism, total utilitarianism, and strong longtermism) and considers technological stagnation to be an existential risk, is likely a rare view among the global population. It may rise in popularity in the future, but presently it appears to be a fringe position.

3.2 The risk of a non-representative view

Tying the study of a topic that fundamentally affects the whole of humanity to a niche belief system championed mainly by an unrepresentative, powerful minority of the world is undemocratic and philosophically tenuous. Landemore defines the term "elite" as a group of people that would not likely be selected at random from its wider population and that is granted decision-making powers.[58] Under this definition, the field of existential risk is decidedly elitist at present. There are ways to mitigate against elitist research projects: diversifying the field and thus its policy recommendations, and democratising the evaluation of policies that are proposed by the field (see Section 6).

An obvious retort here would be that these are scholars, not decision-makers, that any claim of elitism is less relevant if it refers to simple intellectual exploration. This is not the case. Scholars of existential risk, especially those related to the TUA, are rapidly and intentionally growing in influence. To name only one example noted earlier, scholars in the field have already had "existential risks" referenced in a vision-setting report of the UN Secretary General. Toby Ord has been referenced, alongside existential risks, by UK Prime Minister Boris Johnson. Dedicated think-tanks such as the Centre for Long-Term Resilience have been channelling policy advice from prominent existential risk scholars into the UK government.[59]

The field also appears to be elitist in the more common-sense notion of being representative of a small stratum of people with disproportionate economic and political power. The main research centres are clustered in a few of the most elite universities in the world, with most of the field located in Oxford, Cambridge, or the San Francisco Bay Area. As noted earlier, the field is disproportionately supported by billionaires and millionaires. They not only hold financial power, but often advisory positions as well. The ideas of the TUA also closely echo popular Silicon Valley ideology.[60] Indeed, in 1995 Barbrook and Cameron referred to the techno-utopian ideas of the "Californian Ideology", which distinctly echoes the TUA. The Californian Ideology began in Silicon Valley during the tech boom of the 90s and was underpinned by a commitment to technological determinism and neoliberal economics.[61]

The point is a broader one: it is highly risky to grant privileged influence over the fate of *Homo sapiens* to a tiny minority. This is true for the study of existential risk, but more so for the implementation of policies that are meant to reduce existential risk, which will need to balance trade-offs between different interests. An attempt to reduce elitist interference in the study and implementation for existential risk mitigation is important due to moral uncertainty.[62] For every moral theory, there exist recommendations which fail to match our intuitions.[63] It is thus all the more important to know what choices would empirically be preferred by widening the range of people that are allowed to decide what risks are worth and not worth taking.

Some scholars associated with the TUA have written about moral uncertainty. This is excellent and should continue. They advance theories such as expected moral value (ranking alternative axiologies by their expected value,[64] or expected moral choice-worthiness (weighting moral theories by credence and combining them)[65] to navigate moral uncertainty. Issues remain for these approaches, including whether moral theories are comparable in this manner, whether empirical and moral uncertainty are equivalent, how such approaches are scaled up to collectives, as well as technical problems.[66] Using the suggested approaches tends to lead to total utilitarian perspectives outweighing others once large populations in the future are assumed.[67] There is room for considerably more research, and approaches to dealing with moral uncertainty have yet to be consistently and practically applied to existential risk.

3.3 Existential risk: Who is threatened?

The original definition of the techno-utopian paradigm is not concerned with humans *per se*. Instead, it is focused on Earth-originating intelligent life, and enhanced posthumans. This is a different inquiry to studying human extinction, and it is not obvious that it should be conducted under the banner of existential risk. The existing species *Homo sapiens* can be approached empirically, offering the opportunity to develop a science of existential risk; by taking an interest in the future of *Homo* *sapiens*, scholars can approach existential risk reduction as a communal project, able to engage with the subject of inquiry — existing humans — and consider their individual preferences and visions of what a good future would look like.

4. Flawed Definitions, Frameworks and Tools: How Ambiguity and Unverified Assumptions Undermine Research into the Apocalypse

4.1 Ambiguous definitions

This section will look at the problem of defining existential risk. We look at three definitions that can be considered part of the TUA, where an existential risk is one that:

a) "threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development";[68]

b) "threatens the destruction of humanity's long-term potential" (through extinction, unrecovered collapse, or permanent dystopias);[69]

c) "causes the loss of a large fraction of expected value".[70]

The three definitions share a core feature: they are all fixated on future value. The tragedy to be averted is not the suffering or loss of existing humans, but rather the loss of future value or potential. All the definitions are motivated by future, long-term value. There are also some differences. In his extended typology (see Section 2.1) Bostrom's definition enshrines a particular moral view by specifying desirable futures as technological maturity. The other two definitions are, at least in theory, more abstract and value-agnostic.

By leaving "value" and "potential" undefined, these latter definitions theoretically avoid the charge of existential risk as being a project of a niche philosophical view. According to Ord, our potential should be the entire set of possible futures. He suggests that we should first reduce existential risks to a minimum level to achieve "existential security" before undertaking a "Long Reflection": a patient, collective discussion of what exactly humanity's potential is. There are significant problems with this approach.

First, in practice, value is still expressed in techno-utopian terms. For example, the last chapter of *The Precipice* expands on a vision of humanity's potential: transhumanist space expansion receives ample attention and adoration. Unrecoverable civilisational collapse (a state in which technological progress is not ensured) is described as an existential risk. Here, "civilisational collapse" refers to a permanent reversion back to non-agricultural ways of living. It is not explained why the presence of agriculture, or many of the commonly assumed trappings of "civilisation", such as urbanism, writing, and states (although these rarely came as a coherent package; see Graeber and Wengrow)[71] would increase the likelihood of reaching our potential. For a techno-utopian, it does. For others who value virtue, freedom, or equality, it is unclear why a long-term future without industrialisation is abhorrent: it all depends on one's notion of potential. The definition is seemingly agnostic in the abstract, but in practice there are numerous signals that it expresses the same commitment to total utilitarianism and transhumanism.

Secondly, we need to define what our potential is before we can identify threats to it. How else would we know which risks to address? This is an inherent tension within *The Precipice* since we are supposed to achieve existential security before undertaking the Long Reflection. It is difficult to know if we have achieved existential security if we haven't defined what an existential risk is, since we haven't undertaken the Long Reflection to define our potential. A reasonable counter could be that, in theory, there are certain futures that almost no one would like to live in (such as nuclear winter), and that there may be certain risks (for instance, an asteroid strike) that would take lots of plausibly good options off the table. Extinction may indeed be an outcome which we could assume most people would agree we should avoid. Beyond this point of convergence, there may be far more disagreement on what futures are worth protecting.

Indeed, agreement on the badness of extinction and disagreement over our potential is evident in wider philosophical debates. Philosopher Elizabeth Finneron-Burns argues that extinction is wrong due to the suffering and psychological traumas it could cause, not due to the prevention of millions or billions of people yet to be born.[72] Similarly, an intrinsic end value of humanity can be grounds to want to ensure the long-term survival of humanity, but not the potential for many additional future lives.[73] Others have argued that the badness of extinction is a generation-centred issue and neither the future masses or appeals to the natural lifespan and shape of humanity (there isn't one) provide sufficient grounds.[74] There is a strong case that different value theories should be concerned about extinction. There is not a compelling case that many are compatible with longtermist concerns about the deep future.

Third, the "Long Reflection" and Ord's definition assumes we can reduce risks to humanity's potential without choosing between conflicting values. This is almost certainly not the case. Maybe Ord means that we should reach existential security without closing off any possible futures, but retaining option value is presumably restricted to future options that are morally valuable. Indeed, the very notion of existential risk presumes that certain futures, such as dystopias, are to be avoided. We must define what is morally valuable to identify what is dystopian, and we cannot wait until a "Long Reflection" to do so.

Beyond the simpler domain of extinction, things get far murkier. Different plausibly good futures will often involve trade-offs against each other. Steering the world through an age of perils will involve difficult choices. Choices that will frequently have divergent answers depending on one's values. Look no further than the COVID-19 pandemic to see how comparatively smaller crises lead to clashes in values and understandings of what defines a good society. The field should provide clear delineations between risks that were identified as threatening across a broad swathe of value assumptions, and risks that are only threatening given a particular notion of potential. Some scholars may choose to stick with studying extinction risk, rather than trying to specify all possible good futures and the existential threats to them.

Deciding what risks are worth taking and which risks should be taken seriously will need to be a matter of reflection and collective decision-making if it is to respect moral uncertainty and diverse preferences.

Ord writes: "If we steer humanity to a place of safety, we will have time to think",[75] but who is "we"? What appears a risk worth accepting to some will not be considered a risk worth taking by others. For instance, slowing technological progress is a risk to a transhumanist who sees our potential in overcoming human biological limitations, but accelerating technological progress may be perceived as a risk by others. In the following sections, we will list several examples of proposed mitigation efforts which show that judgements of strategies to mitigate risk depend on subjective notions of value. Generally, it is easier to identify universal extinction risks than universal existential risks.

Fourth, in practice the attempt at neutrality can end up masking rather than eliminating values from the analysis. Explicit commitments to transhumanist values in Bostrom's definition of existential risk have the advantage of transparency. It is easier to reject or counterbalance a researcher's perspective when their underlying values are clear. Abstract definitions can end up implicitly incorporating moral assumptions. This can happen unbeknownst to the researcher.

Fifth, the definitions are sufficiently ambiguous to render them unfit for use in a rigorous or replicable risk assessment. If we conceptualise an existential risk as the "permanent and drastic destruction" of desirable future development,[76] human potential,[77] or expected value,[78] then how severe does the destruction need to be? How significant must be the loss of human potential or expected value? How high does the likelihood of its permanency need to be? Even if a risk is judged sufficiently impactful, how large does the probability of the risk occurring need to be, before it can be considered as a legitimate existential risk? How should we compare such risk being incurred by some action against potential benefits of said action? Should risks based on speculation or thought experiments or naive technological extrapolation be treated seriously (see Section 5.2)? These decision-relevant ambiguities have not yet been clarified.

These are not minor, theoretical quibbles. Empirical and moral assumptions determine what is considered an existential risk and what mitigation efforts are recommended. Whether or not these choices are considered reasonable doesn't depend on a replicable framework. Instead, it relies on whether the judge shares the same assumptions.

An example: Ord[79] considers Artificial Intelligence (AI) to be the biggest contributor to existential risk within the next 100 years. The author's probability estimate relies on a survey of machine learning researchers,[80] a study with questionable methodological value for determining whether AI is an existential risk.[81] The policy recommendations for mitigating such risks in *The Precipice* support R&D into aligned Artificial General Intelligence (AGI), instead of delaying, stopping, or democratically controlling AI research and deployment.[82] These policy recommendations were echoed in the "Future Proof" report by the Centre for Long-Term Resilience,[83] which was aimed at UK policymakers. This recommendation seems to assume a kind of technological determinism (see below), or an implicit advocacy for building advanced technology for instrumental purposes. Either explanation echoes the TUA and leads to recommendations that support a particularly existentially risky course of action: developing advanced AI. Despite an explicit acknowledgment that AI could be a major contributor to risk and that slowing, delaying or halting development can help avoid the risk, the book recommends R&D, rather than e.g., citizen surveys, moratoria, or transparency measures.

What appears to be a risk worth accepting to some will appear to be a risk not worth taking to others. Take, for example, Bostrom's "Vulnerable World Hypothesis",[84] which argues for the need for extreme, ubiquitous surveillance and policing systems to mitigate existential threats, and which would run the risk of being co-opted by an authoritarian state.[85] It is a solution that some may find appealing and others appalling. Without a deliberation between different moral views, it should not be assumed that this risk is acceptable. Given disagreements about risk-taking, the field needs to ask who gets to put forward recommendations and who gets to choose between them (see Section 6).

Many definitions face the challenge that they are too abstract to allow for robust, replicable analysis, but the TUA's definitions of existential risk are particularly faulty. These definitions conflate the study of global catastrophe or human extinction with that of the longtermist ethics of existential risk. The question of what futures are worth taking which risks for will always rely heavily on a notion of value. For this reason, we suggest scholars consider separating the following areas:

- *Extinction Ethics*: the study of the ethics of human extinction; the badness or goodness of human extinction given different ethical considerations.
- *Existential*[86] *Ethics*: the study of the ethical implications of different societal forms. This includes not just our potential in the deep future, but also how societies should be structured in the present. This in turn provides ground for defining what would constitute what the field currently calls an *existential* risk. An easier initial step here may be to specify dystopias that most value theories wouldn't want, rather than develop a widely shared notion of potential.
- *Catastrophic Risks*: the study of contributors to the occurrence and probability of global catastrophic events.
- *Extinction Risks*: the study of contributors (which include global catastrophic events) to the probability of human extinction.

While each inquiry can have both empirical and theoretical elements, the latter two lend themselves to scientific risk analyses, while the former two are more philosophical inquiries by nature. All the existing definitions currently conflate all of these. For many analyses, this is unnecessary and counterproductive. The media repeatedly makes the mistake of equating Ord's estimated one in six chance of existential catastrophe with extinction risk,[87] but these are meant to be drastically different concepts.[88] The study of extinction and existential (or longtermist) ethics does not need to be resolved to scientifically study risks, and it requires a different set of skills and procedures (see Section 6). There will be some necessary and fruitful areas of overlap. For instance, ensuring that measures to prevent and mitigate risks are proportional and not dangerous, we will need to have some debate on existential ethics.

Risk analysis will always be at least partly subjective and value-laden. Yet, it can be made more objective and scientific through precise definitions, the transparent statement of assumptions, and where possible, separating risk assessment from the study of extinction and existential ethics.

4.2 Arbitrary categorisations

ERS currently lacks a framework or methodology to categorise risks consistently, comprehensively, and rigorously. The TUA does not provide such a framework. It purports to distinguish between existential and catastrophic risks, but the distinction is hazy and arbitrary.

ERS currently distinguishes between existential risks and "global catastrophic risks". All existential risks are global catastrophic risks, but not all global catastrophic risks are existential risks. Definitions of global catastrophic risks proposed in the literature vary but tend to focus on a significant global loss of human life (such as a loss of 10%).[89] Most include the great catastrophes of the past, such as the Black Death.

Under the earliest TUA definition, a catastrophe that does not jeopardise the attainment of technological maturity is assigned comparatively little moral consideration. This is due to the belief that such disasters have not influenced our long-term fate, and thus do not constitute existential risks,[90] and that limited resources would be wasted if they were directed towards global catastrophes that did not threaten technological maturity.[91] This extreme prioritisation of existential risks downgrades the importance of addressing other global catastrophes. What under Bostrom's view would plausibly be considered "feel-good projects of suboptimal efficacy", and what falls short of "efficient philanthropy",[92] could in fact very much be worth the attention of those who study human extinction.

Ord[93] presents a more moderate version of this original TUA framework. He considers it justified to spend some resources on large catastrophes, because those catastrophes could indirectly amplify (but not directly cause) existential risk. His framework still draws clear distinctions between existential and catastrophic risks framing some as direct “risks” and others as indirect “risk factors”, prioritising the reduction of direct risks.

We lack the necessary understanding of how human society and the Earth system operate to make such neat, surgical distinctions. Whether a single global catastrophe can or would fundamentally alter the trajectory of humanity is one of the great unanswered questions of history. For now, we struggle with forecasting GDP even a year ahead[94] and in fact reliably make incorrect forecasts about population growth, despite knowing all essential variables.[95] In a complex adaptive system it may be impossible to forecast how small changes will affect the longue durée. Neat distinctions between GCRs and existential risks presume a level of systemic knowledge and certainty we currently do not possess. The TUA thus far does not offer the tools to make any fine-grained, credible separation between existential and non-existential global catastrophes. It has not offered explanations of how such events affect extinction and existential risk.

This is not to say that some persistent, long-term, societal trends cannot be identified and understood. For instance, the historian Walter Scheidel has put forward a compelling empirical case that (intra-country) wealth inequality increases inexorably until a great leveller (a state collapse, pandemic, revolution, or mass mobilisation warfare) resets the playing field.[96] Investigating these trends will be critical to foreseeing how catastrophic risks are produced and could unfold deep into the future. Such an analysis does not require or justify the crude split between global catastrophes and extinction risks.

Under the TUA, an existential risk is understood as one with the potential to cause human extinction directly or lead us to fail to reach our future potential, expected value, or technological maturity. This means that what is classified as a prioritised “risk” depends on a threat model that involves considerable speculation about the mechanisms which can result in the death of all humans, their respective likelihoods, and a speculative and morally loaded assessment of what might constitute our inability to reach our potential.

Imagining pathways to human extinction (kill mechanisms) invariably requires some creativity and speculation. This has meant that some areas of risk (e.g. AI) which are not as empirically constrained are often prioritised above others for which we have far more empirical data (e.g. climate change). This has led some (non-peer-reviewed) publications (including a Google Doc) to conclude that climate change is not an existential risk.[97]

This is the wrong question to ask. Adding up the predicted impacts of a selection of hazards and asking "Will the total of these kill everyone?" is a simplistic and ineffective way of conducting risk analysis. This is not how risk unfolds in reality: hazards interact with networks of societal vulnerabilities and responses as well as each other, and can trigger cascading failures.[98] We need to consider different pathways and ways in which climate change (or any other source of risk) can contribute to the overall level of extinction or catastrophic risk we face.[99] The question of "Is this an existential risk?" is naive. We should instead ask: in a given world-state (with structure, vulnerabilities, and the capacity for change) how much will a given process or event increase the overall likelihood of human extinction, and what are the plausible[100] pathways for it to contribute to extinction risk?

A field looking for the *one hazard to kill them all* will end up writing science fiction. More speculative risks are prioritised because a seemingly more complete story can be told and speculative mechanisms by which AI could kill every human can seemingly not yet be ruled out.

In practice this could be addressed by lowering the threshold of what risks should be treated as relevant to extinction and include more of what Ord calls "risk factors", such as those commonly thought of as Global Catastrophic Risks (GCRs). Global catastrophes and responses to GCRs can give vital insight into vulnerabilities that should be mended and resilience factors that should be enhanced.

The distinction between "direct" and "indirect" risk factors also depends on speculation. Direct risks appear to be those for which we can tell a story about how they might "directly" (presumably limited to third or fourth order effects) lead to the extinction of *Homo sapiens*. Unaligned AI, for example, is often considered a direct risk, but the story about extinction from AI is far from complete.[101] Strong expert disagreement regarding risks from AI[102] is testament to how debatable the empirical foundations of "direct" risk pathways of AI still are. Additionally, risks originating from AI could come in many forms, each relying on speculation about different kill mechanisms and assumptions about the nature and use of a system that has never been built. Surely, not all these pathways are equally direct, and yet AI is prioritised as a seemingly homogenous hazard-cluster across key texts within ERS.[103]

A risk perception that depends so strongly on speculation and yet-to-be-verified assumptions will inevitably (to varying degrees) be an expression of researchers' personal preferences, biases, and imagination. If collective resources (such as research funding and public attention) are to be allocated to the highest priority risk, then ERS should attempt to find a more evidence-based, replicable prioritisation procedure.

4.3. Simplistic risk models

4.3.1 Complex vs. crude risk assessments

Risk assessment has evolved dramatically in past decades. Scholars now commonly analyse systemic risk (the ability for a single disruption to cascade into systems failures),[104] how risks can cascade across borders and sectors,[105] and how failures in critical systems can synchronise and reinforce each other.[106] This has led to new forms of complex risk assessment, particularly in climate science and disaster risk reduction. The Intergovernmental Panel on Climate Change (IPCC) sees risk as composed of vulnerabilities, hazards, and exposures, as well as response risks.[107] Similarly, others have suggested that a complex risk assessment needs to consider four determinants of risk (hazard, vulnerability, exposure, response) as well as how risks link and cascade. Understanding the common drivers across each of these determinants is critical to mitigation efforts.[108]

How we assess risk is fundamental to what we consider as an existential or catastrophic risk. For instance, Ord focuses on super volcanoes as a potential existential catastrophe. However, lower magnitude volcanic eruptions could have catastrophic impacts due to their cascading effects and the vulnerable nature of critical infrastructure systems.[109] Similarly, stratospheric aerosol injection[110] does not appear to pose direct risks that would classify it as a global catastrophic threat. Yet, this depends on how it is deployed, the world in which it operates, and the level of warming it is masking. If another calamity, such as a volcanic eruption, solar flare, or nuclear war, destroys the mitigation system for a prolonged period, the ensuing "termination shock" (rapid global warming over a short timeframe) would likely result in catastrophic effects.[111] A more complex risk assessment will be more difficult to do, but it will be more accurate, realistic, and informative.

Most existential risk texts take a simpler, hazard-centric approach. They tend to focus on a few selected hazards: biologically engineered pandemics, Artificial General Intelligence (AGI), nuclear war, climate change, and asteroid strikes.[112] As currently framed, TUA equates risk with hazard and ignores the wider literature on risk assessment in fields such as disaster risk reduction. It is also unclear how the TUA suggests systematically clustering, prioritising, or analysing these hazards: current attempts rely on simplistic categories of "Natural", "Anthropogenic", and "Future" hazards[113] or presenting the selected hazards as the ones worth discussing without explanation. There have been some recent attempts to provide alternative frameworks[114] but these have found little application thus far, and still do not consider response risks and many other relevant areas. Research efforts are often split across the lines of these different hazards. Working across them as part of a more complex risk assessment could offer novel insights.

4.3.2 Technological determinism

The choice to structure risk assessment this way has not been explained or defended. It may have been chosen due to an implicit techno-determinist threat-model: the TUA often appears to assume an exogenous threat model in which existential hazards naturally and apolitically arise from inevitable and near-autonomous technological progress. The TUA rarely examines the drivers of risk generation. Instead, key texts contend that regulating or stopping technological progress is either deeply difficult, undesirable, or outright impossible.[115] Bostrom proposed a "Technological Completion Conjecture": if technological developments do not cease, then all important, basic technological capabilities will be obtained in the long run.[116] Others offer a more sophisticated view, in which military-economic competition exerts a powerful selection pressure on technological development.

This "military-economic adaptionism" constrains sociotechnical change to deterministic paths. Technologies that gift a strong strategic advantage will almost certainly be built.[117] Many in the related Effective Altruism community disregard controlling technology on the grounds of a perceived lack of tractability.[118]

Whether it is technological determinism, the more nuanced military-economic adaptationism model, or concerns around tractability, the result is the same: regressing, relinquishing, or stopping the development of many technologies is often disregarded as a feasible option.

The proposed alternative is "differential technological development": speeding up and slowing down different technologies to ensure they occur in the safest order possible. Why this is more tractable or effective than bans, moratoriums, and other measures has not been fully explained and defended (see Section 5.1 for further discussion). This could be interpreted as hard-nosed pragmatism, an argument that we can stop technologies from being built, but it will not be an efficient or prudent use of resources. Again, a compelling analysis has not been made for why this is the case, and in practice this ends up looking identical to technological determinism. The irony is that if the world is locked into the development of dangerous technologies, then we are already in a "lock-in scenario" so dreaded by many within the TUA.[119] In the eyes of the TUA, the range of future options available to humanity is already greatly restricted.

It is unclear whether technological determinism in the TUA is descriptive or prescriptive. It could be a genuine belief that controlling technology is infeasible. It could also be that under the TUA unabated technological progress is vital to achieve technological maturity and avoid existential risk.

In any case, the assumption of technological determinism leads scholars to focus on hazards, rather than, say, exposure, maybe because there appears to be no point in trying to reduce humanity's exposure to a technology since the development of the technology is assumed inevitable. It is then merely a question of whether benevolent technologies are built first. Attempts to change political or economic drivers of different risks get less attention. This unstated threat model leads to sharp divergence from modern developments in risk analysis towards a crude hazard-centrism.

Importantly, assumptions around technological determinism are highly contested. Indeed, technological determinism is largely (for better or worse) derided and dismissed by scholars of science and technology studies.[120] We have historical evidence for collective action and coordination on technological progress and regress. One example is weather modification. Early attempts were made by the US during the Vietnam War to use weather modification technologies to extend the monsoon season and disrupt enemy supply chain.[121] The introduction of the 1976 Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques (ENMOD Convention) seems to have successfully curtailed further research into the area.[122] The assumptions on technological progress must be thoroughly examined, empirically and theoretically, before they should be used to determine policy and mitigation actions.

There are enough exceptions to doubt that any strategically powerful technology will be due to competition between (largely) rational actors (usually states). Many important technologies such as glass and steam engines were used for ceremonial purposes for centuries before being redirected towards practical purposes.[123] During the early industrial revolution, water mills were more reliable and efficient than coal-fired steam engines. The latter were adopted not because of their inherent superiority, but because they could be located in urban areas with a large and desperate population, which appealed to early capitalists.[124]

Any approach to existential risk will struggle to find frameworks that are comprehensive yet elegant and practical. It will need to be transparent about its empirical assumptions, including on how risks are created. For now, the hazard-centrism and opaque technological determinism of the TUA provides a framework that is overly simplistic, unduly curtails the available mitigation options, and provides no compelling method to understand or address the common drivers behind risk determinants. It is inadequate for the grand challenge of understanding and mitigating extinction risks.

There is room for different empirical worldviews, different frameworks, and different moral positions to be considered in the study of existential risks. We do not at all recommend that hazards should no longer be studied. Similarly, speculation will always be a part of assessing unseen risks, but a science of extinction should adopt frameworks which minimise the need for this aspect. Different frameworks will make different predictions about what policy and research efforts appear to plausibly reduce risk.

4.4 Inappropriate translations of theory into practice

Mitigating existential risk requires decision-making under uncertainty. Decision theory in the context of ERS and longtermism is an active area of research. For now, we want to caution against applying idealised decision-theoretic results to the evaluation of risky choices in practice. This is because empirical uncertainty can affect the applicability of results that hold in theory and because expressing subjective notions of risks and benefits numerically can provide a false sense of certainty.

Take Expected Value (EV), defined as the value of the outcome multiplied by the probability of it occurring. The TUA extensively uses Expected Value calculations to justify its own approach and prioritisations. For example, Bostrom argues that existential risk mitigation should be prioritised over other altruistic acts: if there is just a 1% chance of 10^{54} people coming to exist in the future, then "the expected value of reducing existential risk by a mere one billionth of one billionth of one percentage point is worth a hundred billion times as much as a billion human lives".[125] According to Bostrom, "even the tiniest reduction of existential risk has an expected value greater than that of the definitive provision of any "ordinary" good, such as the direct benefit of saving 1 billion lives".[126] Elsewhere, Millett and Snyder-Beattie use EV to argue for reducing risks from biological pathogens.[127]

While EV is a useful theoretical tool in a range of contexts, in practice, it is hard to apply rigorously when working with the generally low and highly uncertain probabilities characteristic of existential risks. The TUA applies expected value theory to the very areas where it faces the most pitfalls, that is, situations of deep uncertainty and low information about probabilities.[128] Ord attempts to estimate probabilities of existential catastrophes caused by various hazards for communicative purposes, most notably the aforementioned one in six figure.[129] While this has been successful in terms of public communications (the estimate has been widely reported in the press), it is unclear how to evaluate the accuracy of these estimates or whether his methodology for arriving at them is sufficient to warrant the sense of scientific credibility that numbers inevitably imply to the lay person. In addition, to evaluate the EV of mitigating an event, one must decide upon a particular conception of value. Thus, personal intuitions and non-representative moral preferences are at risk of being captured and made to appear objective by numbers. How can policy recommendations be considered robustly good and replicable if they are evaluated on subjective assessments of probabilities and hidden ethical assumptions? EV is still unsuitable to be relied upon in practice for estimating human impact on the value of the long-term future.[130]

Furthermore, EV and decision theories more widely are affected by Pascal's Mugging,[131] as well as what has been called *fanaticism.*[132] We know of no pragmatic and consistent response to those challenges yet. Fanaticism describes how we may be required to put considerable effort into mitigating terrible events with an arbitrarily miniscule probability of occurring. Similarly, Pascal's Mugging describes how vast or near-infinite quantities of value can overwhelm even the most minuscule probabilities: the term comes from a thought-experiment in which Pascal is conned by a self-proclaimed wizard who promises to magically grant 1,000 quadrillion happy days in exchange for his wallet.[133] The probability that the grifter is a powerful wizard is infinitesimally low, but not zero, and outweighed by the expected utility of so many happy days. The Pascal's Mugging problem arises when applying EV to existential risks as defined within the TUA. Any risk that could prevent technological maturity, no matter how small, should be taken seriously and acted on due to the sheer amount of expected value at stake. Scholars have suggested that Pascal's mugger returns to swindle another unsuspecting victim, but this time using existential risk studies and longtermism.[134]

Both these challenges are usually evaded by claiming that hazards like AGI have an unambiguously high enough probability of occurring this century to merit considerable action.[135] This only side-steps the question in the case of AGI in particular (assuming that these doubtful estimates can be trusted) and does not tell us whether highly unlikely or speculative risks should be acted on. There are good reasons to defend fanaticism[136] and further theoretical work might resolve the challenges presented here, but while the integration of theory and empirical work is still ongoing, we should consider drawing pragmatic lines between mere speculation and risks humanity should significantly focus on. For instance, a simple threshold or plausibility assessment[137] could protect the field's resources and attention from being directed towards highly improbable or fictional events.

5. The Risks of Studying Existential Risk

How could the study of global risks and longtermism contribute to catastrophe? The worst-case outcome is not that existential risk remains unaffected, or that resources are wasted on incorrect speculations (although these are problems). Instead, it is that these risks are aggravated by research into them. This issue must be addressed for any approach to ERS. Unfortunately, the TUA appears to be particularly prone to both ignoring response risks and aggravating them.

5.1 A risky road to safety

Mitigating risks incurs a risk of its own. As noted earlier, this is a fundamental part of sophisticated risk assessments, including those used by the IPCC.[138] Not all approaches incur the same risks. While we will not compare the techno-utopian approach against alternatives in this chapter, we think the TUA is prone to an especially high level of response risk.

The zealous pursuit of technological development, according to proponents of the TUA, accounts for the vast majority of risk over the coming centuries. The risk of human extinction from natural hazards is likely low, with an upper-bound of less than one in 14,000[139] and a best guess of around 0–0.05% per century.[140] In contrast, several scholars of existential risk place the likelihood of an existential catastrophe far higher at 1/6,[141] or >1/4[142] over the coming century. Rees puts the chance of collapse or extinction by 2100 at 1/2.[143] This discrepancy is mainly due to anthropogenic risks arising from climate change, nuclear weapons, synthetic biology, and artificial intelligence (AI).

If the lion's share of extinction risk stems from emerging technologies, why do we rarely ask how to stop dangerous developments? This option is usually considered infeasible,[144] or outright impossible.[145] This may be due to the exogenous threat model and technological determinism of the TUA. Since halting the technological juggernaut is considered impossible, an approach of differential technological development is advocated.[146] This involves trying to develop beneficial and protective technologies (aligned Artificial Intelligence[147] is often used as an example) first, before proceeding to riskier options.

It is not clear how scholars plan to reliably determine which non-existing technologies will be more or less risky years or decades in advance. Even if we did have such a refined vision of the future, it is unclear why a precise slowing and speeding up of different technologies (which are interlinked and presumably require a set of fine-tuned regulatory tools) across the world is more feasible or effective than the simpler approach of outright bans and moratoriums.

More importantly, in the TUA the stark choice between one of only two destinies — technological maturity or existential catastrophe — is a *fait accompli*. The path to techno-utopia appears to be the only one available, despite its risks. From a techno-utopian perspective, a failure to build these dangerous, powerful technologies is an existential risk. Bostrom, aware of the tension arising from recommending the (albeit careful) development of technologies, warns: "We should not blame civilization or technology for imposing big existential risks. Because of the way we have defined existential risks, a failure to develop technological civilization would imply that we had fallen victims of an existential disaster. [...] Without technology, our chances of avoiding existential risk would therefore be nil".[148] The TUA dramatically restricts the options available for avoiding existential catastrophe.

Furthermore, pursuing a techno-utopian future is dangerous and may come with considerable cost. It has already been noted that the attempt at colonising space and an expansion of technological capabilities could end in catastrophe if it foments a new arms race and large-scale warfare.[149] Upgrading the human body could construct a biological caste system, where, an enhanced, genetic elite could oversee a subjugated, unenhanced, "inferior" class.[150] These are not far-flung speculations. Researchers already actively monitor, evaluate and debate near-term bio-engineering enhancements and their ethical implications.[151] Enhanced inequality would not only be unjust, but also amplify many social ills.[152] Similarly, a horizon scan under the WHO Science and Research Division has noted that the pursuit of advances in the life-sciences could produce many technologies that could be easily misused to cause harm. These range from using bioregulators for the delivery of bioweapons to the use of deep learning algorithms to identify novel biological pathogens.[153]

Existential risk does not need to be defined in reference to technological maturity, nor does it need to be accompanied by these response risks of accepting or even speeding up disruptive technologies. A different vision for a good future could lead to dramatically different policy recommendations. A less determinist view of technological change and pessimistic view of political change would open up a plethora of other interventions. A democratic approach to ERS will provide ample room for different moral and empirical assumptions to affect the assessment, discussion, and negotiation of collective risk-taking. Other paths and approaches may be more risk-averse and must be explored if humanity wants to safely reduce existential risk.

5.2 High stakes: Existential exceptions and a risk-averse approach to existential risk

5.2.1 The Stomp Reflex

There is a long history of security threats being used to enable draconian emergency powers. Emergency powers are intended to be conservative: to protect existing legal and political structures in a period of tumult. The logic is that drastic times call for drastic measures. To protect institutions, emergency powers allow governments to disregard existing laws and exempt themselves from judicial or democratic restrictions and oversight. Rather than protect, such measures are often abused to erode and transform fundamental political structures; when trying to centralise and extend state powers, fear is a powerful justification. The larger the fear, the easier it is to justify more potent emergency powers. If the perceived threat is human extinction, then the measures could be extreme.

Recent examples abound. The Patriot Act, adopted by the US just 45 days after the 9/11 attack, allowed for a range of draconian actions, including indefinite detainment of migrants without criminal prosecution. The provisions broadly underpin the current US surveillance network, most notably through the NSA's PRISM program, which Edward Snowden exposed in 2013. The *War on Terror* became a useful cover for the creeping power of the US security apparatus. This despotic drift is not a purely historical threat. Clauses for states of emergency have spread over past decades[154] and 2020 marked the highest ever use of such measures.

The prolific use of emergency powers can lead to the creation of a *state of exception* in which the sovereign transcends the regular rule of law. Temporary measures become permanent, and spill into the operation of the legal system.[155] The transition from the Roman Republic to the Roman Empire, the fall of the Weimar Republic in the 1920s and 1930s into the Nazi regime, and many other political declines were underpinned by the normalisation of emergency powers.[156]

The irony is that emergency powers rely on an inaccurate understanding of human nature and disaster risk. Emergency powers inevitably empower those atop hierarchies, despite abundant evidence from disaster risk reduction and other fields that while mass panic is a myth, the risk of elite panic and elite co-option of catastrophes is real.[157] There is even evidence that such a response worsens crisis. One study of natural disasters found that the larger the number of emergency provisions used by an executive, the higher the fatalities (controlling for disaster severity and size).[158] This is a "*Stomp Reflex*": governments using emergency powers to reassert and veil systems of authority. Such a response is counter-productive and ultimately shifts power into the shadows, away from transparency and public accountability.[159]

Existential risk is the perfect excuse for enacting the Stomp Reflex. Indeed, catastrophic hazards such as nuclear weapons have already been used to justify anti-democratic shifts. In the US, the accumulated nuclear stockpile and threat of sudden war justified a profoundly autocratic move: a single individual — the President — was given the ability to launch nuclear attacks. Richard Nixon once boasted "I can go into my office and pick up the telephone and in twenty-five minutes seventy million people will be dead". As sociologist Elaine Scarry has argued, the nuclear decision-making apparatus violates constitutional rights, the deliberative nature of democracy, and any social contract. The world lives in the shadow of a "thermonuclear monarchy" rather than a democracy.[160]

Nuclear weapons also saw a revolution in secrecy in the US. The threat of thermonuclear war was used as a justification to construct unprecedented levels of secrecy in the military and intelligence communities. These were of dubious efficacy in preventing the spread of nuclear weapons, but they did have the effect of eroding transparency and democratic control over the military industrial complex.[161]

The best empirical example of policy responses to an existential threat do not inspire confidence. In the US at least, the threat of thermonuclear war spurred dramatic reforms that made for a less democratic and open state, but not necessarily a safer one.

5.2.2 Survival through security, surveillance and suppression

Any approach to understanding and mitigating existential risks runs the risk of becoming *securitised*. Securitisation refers to a discursive manoeuvre that moves an issue from the arena of normal politics to that of national security, making it more likely to permit emergency powers and be placed under the control of unelected military and intelligence officials. Moves towards thermonuclear monarchy and elevated secrecy were largely underpinned by neorealist foreign policy and game theory developed in such a context.[162] This is not to say that all securitisation approaches are equally dangerous; some are far more likely to enable authoritarian responses.

There are reasons to expect that the TUA is particularly vulnerable to misuse. As philosophers such as Peter Singer and Phil Torres have noted, if the world is viewed from the TUA's lens of existential risk, then we run the risk that almost any action is justified if it is believed to improve our chance of surviving to expand beyond Earth.[163] Problems which are not considered to be an existential risk dwindle into irrelevance, as other values are sacrificed on the altar of expected astronomical value.

This is not to say that most believers of the TUA are intent on using it to justify morally abhorrent actions, nor that they are unaware of these weaknesses. Rather, we argue that the basic logic of protecting a high-tech future of astronomical value could be easily co-opted. Marx never intended for communism to justify brutal dictatorships. Nonetheless, it was easily twisted by Stalin and others to do so.

Scholars of existential risk have already shown some proclivity for invoking security, whether it be for "existential security"[164] or "epistemic security".[165] Extreme emergency responses have also been raised. Bostrom's "Vulnerable World Hypothesis"[166] identified the combination of ubiquitous surveillance, preventative policing, and global governance (understood to be "a world government; a sufficiently powerful hegemon or a highly robust system of inter-state cooperation")[167] as a comprehensive agenda to protect against possible technological hazards that could devastate civilisation. He proposes a typology of four potential threats: "easy nukes" (readily accessible and easy to use weapons of mass destruction), "safe first strike" (the ability to safely destroy others with impunity), "surprising strangelets" (experiments that could harbour an unforeseen, or foreseen but low-probability catastrophe); and those in which the accumulation of minor damages by individuals eventually accumulate into global catastrophe.

Bostrom's preferred solution — extreme preventative policing and widespread surveillance — could involve the mandatory use of ironically-named "freedom tags" fitted with multiple cameras and microphones to continuously track individual behaviour. These would be distributed to all citizens and monitored by state employees — "freedom officers" who themselves are watched by artificial intelligence to prevent misuse — who can order preventative interventions using drones or police.[168]

Both the journal paper (published in the journal *Global Policy*), as well as public-facing spin-off articles[169] about the Vulnerable World Hypothesis feature clear policy recommendations. A box in the 2019 paper titled "Policy Implications" includes the recommendations that dealing with "black balls" (technological innovations that by default destroy the world) would require "a system of ubiquitous real-time worldwide surveillance. In some scenarios, such a system would need to be in place before the technology is invented".[170] The public-facing article from 2021 asks: "If you find yourself in a position to influence the macroparameters of preventive policing or global governance, you should consider that fundamental changes in those domains might be the only way to stabilise our civilisation against emerging technological vulnerabilities". It is not difficult to foresee how such ideas could provide grounds for aspiring autocrats to subvert democratic institutions in the face of global threats.

Bostrom[171] also includes a discussion of pre-emptive strikes. He describes the responsibility and need for nations to (on some occasions) enact pre-emptive, unilateral infringements of sovereignty. If extinction threatening technologies (he imagines biosphere-destroying nanobots) are not controlled under international treaties, "the mere decision to go forward with development of the hazardous technology [...] must be interpreted as an act of aggression" and would justify pre-emptive infringement of national sovereignty.[172]

There is a clear danger in authoritative recommendations based on speculative thought experiments. Scholars using the TUA providing recommendations for surveillance and pre-emptive measures in the name of avoiding catastrophe could contribute to birthing the very dystopias they fear.

There is little evidence that the push for more intrusive and draconian policies to stop existential risk is either necessary or effective. It is empirically dubious to think that we cannot halt or delay the development or spread of dangerous technologies. Nor is it convincing that surveillance measures would prove effective. We have little to no evidence that the use of mass surveillance has been effective at preventing terrorist attacks in the US.[173] Moreover, the main creators of such hazards — the *Agents of Doom* — are often the very people who control the surveillance apparatus: military industrial complexes, enormous technology firms, and powerful states.[174] At worst, the knee-jerk reaction of surveillance and preventative policing to prevent a speculative calamity could simply create one of its own: entrenched authoritarianism.

The obvious option to discontinue certain technological developments — if we assume that the Vulnerable World Hypothesis is true — is considered "hardly realistic" and "extremely costly, to the point of constituting a catastrophe in its own right".[175] Ord, too, warns of so-called "desired dystopias", in which an ideology (or manipulation and surveillance) has corrupted our choices to the extent that, for example, we "completely renounce further technological progress".[176]

Under the TUA we appear to be trapped. We either develop technologies which carry immense risk, such that ubiquitous surveillance becomes necessary, or we cease technological development only to manifest the existential risk of failing to reach a technologically mature utopia. This trap is an idiosyncratic feature of the TUA.

There are more options for humanity than merely picking between two highly risky paths. Ord confidently asserts that ceasing any further technological development would "ensure our destruction at the hands of natural risks",[177] but we have seen no convincing analysis that shows we could not safeguard our survival with current technologies and re-directed resources. Indeed, it is unclear why we cannot develop technology to address threats from asteroids and supervolcanoes without indulging in the entire range of dangerous inventions. The opinion that all technological progress must continue at all costs is a dangerous one.

Scholars of existential risk need to be vigilant to response risks, and risk-averse in their own suggested interventions. This means we must be truthful about uncertainty, consider the worst-case outcomes of our actions, and verify the acceptability of proposed interventions by subjecting them to democratic oversight.

6. Democratising Risk

There is an intimate and neglected relationship between existential risk and democracy. Democracy must be central to efforts to prevent and mitigate catastrophic risks. It is also an antidote to many of the problems manifest in the TUA. Do those who study the future of humanity have good grounds to ignore the visions, desires, and values of the very people whose future they are trying to protect? Choosing which risks to take must be a democratic endeavour.

We understand democracy here in accordance with Landemore as the rule of the cognitively diverse many who are entitled to equal decision-making power and partake in a democratic procedure that includes both a deliberative element and one of preference aggregation (such as majority voting).[178] Decision-making procedures are not either democratic or non-democratic, but instead lie on a spectrum. They can be more or less democratic, inclusive, and diverse.

We posit three reasons for why we should democratise research and decision-making in existential risk: the nature of collective decision-making about human futures, the superiority of democratic reason, and democratic fail-safe mechanisms.

Avoiding human extinction, or crafting a desirable long-term future, is a communal project. Scholars of existential risk who take an interest in the future of *Homo sapiens* are choosing to consider the species in its entirety. If certain views are excluded, the arguments for doing so must be compelling.

Democracy will improve our judgments in both the governance and the study of existential risks. Asking how our actions today influence the long-term future is one of the most difficult intellectual tasks to unravel, and if there is a right path, democratic procedures will have the best shot at finding it. Hong and Page[179] demonstrate both theoretically and computationally that a diverse group of problem-solving agents will show greater accuracy than a less diverse group, even if the individual members of the diverse group were each less accurate. Accuracy gains from diversity trump the gains from improving individual accuracy. Landemore[180] builds on this work to advance a probabilistic argument that inclusive democracies will, in expectation, make epistemically superior choices to oligarchies or even the wise few. This is supported by promising results in inclusive, deliberative democratic experiments from around the world.[181] In the long run, democracies should commit fewer mistakes than alternative decision-making procedures. If this is true, it should improve the accuracy of research efforts and decision-making. We are more likely to make accurate predictions about the mechanisms of extinction, probable futures, and risk prevention if the field invites cognitive diversity, builds flat institutional structures, and avoids conflicts of interest.

There are many ways to consider the interests of the many. Democratic assemblies could allow global citizens to deliberate about the futures they prefer, citizens could be surveyed, and the field of ERS itself could be diversified. At the moment, the field is, as many academic disciplines are, unrepresentative of humanity at large and variably homogenous in respect to income, class, ideology, age, ethnicity, gender, nationality, religion, and professional background. The latter issue is particularly true of existential risk, which, despite being an inherently interdisciplinary endeavour, is at the highest levels dominated by analytic moral philosophers. We need to be vigilant to what perspectives are not represented in the study of existential risk. An awareness of bias will go some way towards mitigating its negative effects. To get close to replicating the cognitive diversity found among humans, we must begin by inviting different thinkers with different values and beliefs into the field.

Democracies can limit harms. Any approach to mitigating existential threats could create response risks, and the TUA seems particularly vulnerable to this. Despite good intentions and curiosity-driven research, it could justify violence, dangerous technological developments, or drastically constrain freedom in favour of (perceived) security. If we hope to explore ideas but minimise harms, democracies can be used to moderate the measures taken in response to harmful ideas. It seems, for example, vanishingly unlikely that a diverse group of thinkers or even ordinary citizens would entertain the idea of sacrificing one billion living, breathing beings for an infinitesimal improvement in reaching an intergalactic techno-utopia. In contrast, the TUA could recommend this trade-off.

The democratic constraint of extreme measures may simply be a form of collective self-interest. Voters are unlikely to tolerate global catastrophic risks (GCRs), which incur the death of a sizeable portion of the electorate, if they know they themselves could be affected. We expect that scholars who do not support sacrificing current lives in the name of abstract calculations, but would still like to explore the use of expected value theory in existential risk, will be in support of democratic fail-safe mechanisms.

Empirically, this fail-safe mechanism seems to work. Even deeply imperfect democracies, like the ones we inhabit now, often avert detrimental outcomes. Democracies prevent famines[182] (although not malnutrition).[183] They make war — a significant driver of GCRs — less likely.[184] The inclusion of diverse preferences in democracies, such as those achieved through women's suffrage, further decreases the likelihood of violent conflict.[185] Citizens often show a significant risk aversion in comparison to their government. While surveys are notoriously difficult to collect and interpret, existing data suggest that the public has little support for nuclear weapons use,[186] but strong support for action against climate catastrophe.[187] We can further show that when citizens deliberately engage with the subject at hand, their concern and readiness for action often increases.[188] For example, citizen assemblies on climate change have recommended widespread policy-changes across sectors, amendments to incentive structures and laws against ecocide to reach emissions targets.[189] Indeed, many lament that when it comes to genetically modified organisms and nuclear power, citizens are far too risk-averse.[190] The problem is not that the public is riddled with cognitive biases that make them unconcerned about global catastrophes.

Democratic debate cannot be an afterthought. Navigating humanity through crises will involve many value-laden decisions under deep uncertainty. Democratic procedures can deal with such hard choices. Greater cognitive diversity should be represented amongst scholars of ERS. Recommendations on policies that would reduce risk should be passed through deliberative assemblies and await the approval of a wider pool of ordinary citizens, as they will be the ones who will bear this risk. A homogenous group of experts attempting to directly influence powerful decision-makers is not a fair or safe way of traversing the precipice.

7. Conclusion

The case for the importance of studying existential risk has been made. ERS must now converge on a trustworthy methodology.

The general problems in ERS we identify are that (i) an inclusive definition of existential or extinction risk will require some ambiguity, (ii) any categorisation of risks will be at least partly arbitrary, (iii) any risk assessment will not be entirely comprehensive, and (iv) any study of existential risks and proposed interventions could also increase risk. This field of study is inseparable from a moral inquiry. Definitions of what are catastrophic or desirable futures are inextricably intertwined with questions of value. Dealing with risk is not restricted to risk analysis but includes the question of what risks are worth taking. We believe these challenges are not insurmountable and are worthy of our attention.

The original and influential Techno-Utopian Approach (the TUA) faces specific, daunting problems. These include idiosyncratic and non-representative moral visions of the future, and the use of definitions that are excessively ambiguous and founded on opaque, questionable assumptions. The categorisation problem is exacerbated by combining the study of catastrophe with longtermist ethics. The frameworks are crude and do not include recent advances in risk assessment, or even basic knowledge from directly relevant fields. The TUA does not yet consider response risks and at times advocates for risky policies. It is susceptible to being misused to justify exceptional emergency actions.

We suggest some initial, modest steps for improving the field. First, extinction risks should be analysed separately from extinction ethics and existential ethics. Some research questions will still require combining these inquiries, but the attempt to separate the science of risk from the moral evaluation of risk will benefit each endeavour. Second, existential risk scholars should transparently acknowledge the moral and empirical assumptions used in risk analyses. Third, we must critically embrace the latest advances in risk assessment from other fields, such as climate science and disaster risk reduction. Fourth, and most importantly, existential risk must be cognitively diversified, and the judgement of its recommendations democratised. We can't afford to wait for a "Long

Reflection". Open democracy and collective deliberation need to be central to reducing existential risk and navigating the future.

We encourage existing scholars to enrich the diversity of available frameworks by revising or abandoning the TUA. We encourage researchers to find entirely new approaches and take a more inclusive, participatory approach to thinking about and shaping our responses to potential catastrophes. To have good judgement, represent the interests of the vulnerable, and avoid dangers, the study of existential risk needs to be democratised.

Acknowledgements

We would like to thank the almost two dozen reviewers who took the time to peruse this draft and provide invaluable suggestions. The chapter is much stronger due to your help.

Notes and References

1 Shackelford, G. E. et al. 'Accumulating evidence using crowdsourcing and machine learning: A living bibliography about existential risk and global catastrophic risk', *Futures, 116* (February 2020): 102508. https://doi.org/10.1016/j.futures.2019.102508

2 Rees, M. J. *Our Final Century: Will Civilisation Survive the Twenty-First Century*? Arrow (2004); Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (2016); Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Penguin Books (2018); Russell, S. J. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books (2020); Ord, T. *The Precipice: Existential Risk and the Future of Humanity*. Hachette Books (2020); Christian, B. *The Alignment Problem: How Can Machines Learn Human Values*? Atlantic Books (2021).

3 UN. *Our Common Agenda — Report of the Secretary-General*. United Nations (2021).

4 The following is not a comprehensive summary, but an indicator of the origin of resources in the field: The Future of Life Institute began with a $10 million donation from billionaire Elon Musk. It launched a $25 million program on AI safety with financial support from billionaire Vitalik Buterin. The technology millionaire Jaan Tallinn was a co-founder of CSER, while Elon Musk sits on the Advisory Board. The Future of Humanity Institute has received funding from Elon Musk, as well as over $14 million in grants from Open Philanthropy. Open Philanthropy, founded by billionaires Cari Tuna and Dustin Moskovitz (a cofounder of Facebook), funded the establishment of the Centre for Security and Emerging Technology (CSET) with a $55 million dollar donation. Open Philanthropy has, to date, given over $37 million to projects related to global catastrophic risks, $99 million to biosecurity and pandemic preparedness, and $196 million worth of grants for potential risks from AI (not all of which are related to existential risk) (calculated on the 01/12/2021 using Open Philanthropies online database).

5 Greaves, H. et al., *A Research Agenda for The Global Priorities Institute*. Global Priorities Institute (2019).

6 Ord (2020); Matheny, J. G., 'Reducing the risk of human extinction', *Risk Analysis* 27(5) (7 December 2007): 1335–44. https://doi.org/10.1111/j.1539-6924.2007.00960; Bostrom, Nick. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (February 2013): 15–31. https://doi.org/10.1111/1758-5899.12002; Bostrom, Nick. 'Astronomical waste: The opportunity cost of Delayed technological development', *Utilitas, 15*(3) (November 2003): 308–14. https://doi.org/10.1017/S0953820800004076; Bostrom, Nick (ed.). 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology*, 9 (2002).

7 Bostrom, Nick. 'Letter from Utopia', *Studies in Ethics, Law, and Technology*, 2(1) (9 January 2008). https://doi.org/10.2202/1941-6008.1025; Bostrom, Nick and Philosophy Documentation Center. 'Transhumanist values', *Journal of Philosophical Research, 30*(9999) (2005): 3–14. https://doi.org/10.5840/jpr_2005_26

8 Ord (2020); Bostrom (2013); Bostrom (2002).

9 Greaves, Hilary and Will MacAskill. *The Case for Strong Longtermism*. Global Priorities Institute (2019).

10 Bostrom (2013); Bostrom, N. 'The vulnerable world hypothesis', *Global Policy, 10*(4) (November 2019): 455–76. https://doi.org/10.1111/1758-5899.12718

11 The failure modes we describe in our chapter will thus apply to the elements of the TUA which different publications lean on.

12 Avin, Shahar et al. 'Classifying global catastrophic risks', *Futures, 102* (September 2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001; Liu, Hin-Yan, Kristian Cedervall Lauta and Matthijs Michiel Maas. 'Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research', *Futures, 102* (September 2018): 6–19. https://doi.org/10.1016/j.futures.2018.04.009

13 Liu, Lauta and Maas (2018) suggest examining GCRs in terms of their hazards (what harms humanity), vulnerabilities (how and why a given hazard could cause humanity to come to harm) and exposure (how and why humanity is in harm's way). Avin et al. (2018) classify GCRs by the critical system under threat, the mechanism for spreading risk globally, and failures to mitigate or prevent the risk. See Beard, SJ et al. 'Assessing climate change's contribution to global catastrophic risk', *Futures, 127* (March 2021): 102673. https://doi.org/10.1016/j.futures.2020.102673 for a brief summary, as well as a survey of existing definitions of GCRs.

14 Beard, Simon and Phil Torres. 'Ripples on the great sea of life: A brief history of existential risk studies', *SSRN Electronic Journal* (2020). https://doi.org/10.2139/ssrn.3730000

15 Bostrom (2002).

16 Bostrom (2013).

17 Calculated with Google Scholar on the 10th May, 2021. As an imperfect comparison, one of the few papers proposing an alternative way of thinking about existential risks (Liu, Lauta and Maas, 2018) has a total of 24 citations (calculated on the 10th May, 2021).

18 Bostrom (2003).

19 Bostrom (2016).

20 Ord (2020).

21 Tegmark (2017).

22 MacAskill, Will. *What We Owe the Future*. Basic Books (2022).

23 Kurzweil, Ray. *The Singularity Is Near: When Humans Transcend Biology*. Duckworth (2018); Chalmers, David J. 'The singularity: A philosophical analysis', *Journal of Consciousness Studies, 17*(9–10) (2010): 7–65.

24 Beckstead, Nick. *On the Overwhelming Importance of Shaping the Far Future*. Rutgers University (2013).

25 Karnofsky, Holden. *'Most Important Century' Series: Roadmap* (2021). https://www.cold-takes.com/most-important-century-series-roadmap/

26 A Platonic ideal of a phenomenon that abstracts its essential features, but in reality is observed in many variations.

27 Milanović, Branko. *Capitalism, Alone: The Future of the System That Rules the World*. The Belknap Press of Harvard University Press (2021).

28 Bostrom (2002).

29 Bostrom (2013).

30 Bostrom.

31 Cotton-Barrat, Owen and Toby Ord. 'Existential risk and existential hope: Definitions', *Future of Humanity Institute* (n.d.).

32 Ord (2020).

33 Bostrom and Philosophy Documentation Center (2003).

34 Bostrom (2013).

35 Bostrom (2008).

36 Bostrom and Philosophy Documentation Center (2003); Bostrom, Nick. *What Is Transhumanism*? (2001). https://www.nickbostrom.com/old/transhumanism.html

37 Bostrom (2008).

38 Bostrom (2013).

39 Parfit, Derek. *Reasons and Persons*. Oxford University Press (1986). https://doi.org/10.1093/019824908X.001.0001

40 Matheny, J. G. 'Reducing the risk of human extinction', *Risk Analysis, 27*(5) (October 2007): 1335–44. https://doi.org/10.1111/j.1539-6924.2007.00960.x

41 Bostrom (2013).

42 Bostrom (2014).

43 Sandberg, Anders, Stuart Armstrong and Milan M. Cirković. 'That is not dead which can eternal lie: The aestivation hypothesis for resolving Fermi's Paradox', *ArXiv:1705.03394* [*Physics*] (27 April 2017). http://arxiv.org/abs/1705.03394

44 Brown-Weiss, Edith. *In Fairness to Future Generations: International Law, Common Patrimony, and Intergenerational Equity*. Transnational Publishers (1989).

45 Greaves, Hilary and Will MacAskill. 'The case for strong longtermism', *GPI Working Paper, 5*. University of Oxford, Global Priorities Institute (2019).

46 Tarsney, Christian. 'The epistemic challenge to longtermism', *GPI Working Paper, 10*. University of Oxford, Global Priorities Institute (2019).

47 Bostrom (2013); Tarsney (2019).

48 Ord (2020).

49 Riedener, Stefan. 'Existential risks from a Thomist Christian perspective', *GPI Working Paper, 1*. University of Oxford, Global Priorities Institute (2021); Tarsney, Christian and Teruji Thomas. 'Non-additive axiologies in large worlds', *GPI Working Paper, 9*. University of Oxford, Global Priorities Institute (2020); Greaves, Hilary and Toby Ord. 'Moral uncertainty about population axiology', *Journal of Ethics and Social Philosophy, 12*(2) (2 October 2017): 135–67. https://doi.org/10.26556/jesp.v12i2.223

50 Greaves and MacAskill (2019).

51 Ord (2020).

52 Finneron-Burns, Elizabeth. 'What's wrong with human extinction?', *Canadian Journal of Philosophy, 47*, (2–3) (2017): 327–43. https://doi.org/10.1080/00455091.2016.1278150; Frick, Johann. 'On the survival of humanity', *Canadian Journal of Philosophy, 47*(2–3) (2017): 344–67. https://doi.org/10.1080/00455091.2017.1301764; Lenman, James. 'On becoming extinct', *Pacific Philosophical Quarterly, 83*(3) (September 2002): 253–69. https://doi.org/10.1111/1468-0114.00150

53 Bourget, David and David J. Chalmers. 'What do philosophers believe?', *Philosophical Studies, 170*(3) (September 2014): 465–500. https://doi.org/10.1007/s11098-013-0259-7

54 Hughes, James. *Report on the 2007 Interests and Beliefs Survey of the Members of the World Transhumanist Association*. World Transhumanist Association (2008).

55 Torres, Phil. *Were the Great Tragedies of History "Mere Ripples"? The Case Against Longtermism* (2020). https://www.xriskology.com/mini-book

56 Caviola, Lucius et al. 'Population ethical intuitions', *Cognition, 218* (January 2022): 104941. https://doi.org/10.1016/j.cognition.2021.104941

57 Schubert, Stefan Lucius Caviola and Nadira S. Faber. 'The psychology of existential risk: Moral judgments about human extinction', *Scientific Reports, 9*(1) (December 2019): 15100. https://doi.org/10.1038/s41598-019-50145-9

58 Landemore, Hélène. *Open Democracy: Reinventing Popular Rule for the Twenty-First Century*. Princeton University Press (2020).

59 Ord, Toby, Angus Mercer, and Sophie Dannreuther. *Future Proof: The Opportunity to Transform the UK's Resilience to Extreme Risks*. Centre for Long-Term Resilience (2021). https://www.longtermresilience.org/futureproof

60 Schuster, Joshua and Derek Woods. *Calamity Theory: Three Critiques of Existential Risk*. University of Minnesota Press (2021); Stoecker, Christian. 'Ist "Longtermism" Die Rettung — Oder Eine Gefahr?', *Spiegel Wissenschaft* (2021). https://www.spiegel.de/wissenschaft/longtermism-was-ist-das-rettung-oder-gefahr-kolumne-a-983e60ba-6265-40a8-8c65-8f2668e4e9ff

61 Barbrook, Richard and Andy Cameron. 'The Californian ideology', *Mute Magazine* (1995). https://www.metamute.org/editorial/articles/californian-ideology

62 MacAskill, Will, Krister Bykvist and Toby Ord (eds.). *Moral Uncertainty*. Oxford University Press (2020).

63 Ord (2020).

64 Greaves and Ord (2017).

65 MacAskill, Bykvist and Ord (2020).

66 Kernohan, Andrew. 'Descriptive uncertainty and maximizing expected choice-worthiness', *Ethical Theory and Moral Practice, 24*(1) (March 2021): 197–211. https://doi.org/10.1007/s10677-020-10139-3

67 Greaves and Ord (2017).

68 Bostrom (2013).

69 Ord (2020).

70 Cotton-Barrat and Ord (2015).

71 Graeber, David and David Wengrow. *The Dawn of Everything: A New History of Humanity*. Allen Lane (2021).

72 Finneron-Burns (2017).

73 Frick (2018).

74 Lenman (2002).

75 Ord (2020).

76 Bostrom (2013); Bostrom (2002).

77 Ord (2020).

78 Cotton-Barrat and Ord (2015).

79 Ord (2020).

80 Grace, Katja et al. 'Viewpoint: When will AI exceed human performance? Evidence from AI experts', *Journal of Artificial Intelligence Research, 62* (31 July 2018): 729–54. https://doi.org/10.1613/jair.1.11222

81 Cremer, Carla Zoe. 'Deep limitations? Examining expert disagreement over deep learning', *Progress in Artificial Intelligence, 10*(4) (December 2021): 449–64. https://doi.org/10.1007/s13748-021-00239-1; Cremer, Carla Zoe and Jess Whittlestone, 'Canaries in technology mines: Warning signs of transformative progress in AI', *1st International Workshop on Evaluating Progress in Artificial Intelligence* (24 September 2020). https://doi.org/10.17863/CAM.57790

82 Note this unwillingness to stop or delay progress in technology or research is interestingly often not observed in biorisk research in ERS. It is unclear why this is the case. One explanation could be that banning or stopping technologies and practices in biorisk is seen to be more feasible than doing the same for other areas. We have yet to see a convincing argument as to why this should be the case. Another explanation could be that this is about desire, not tractability. AI is central to the techno-utopian vision of the future. Gain-of-function experimentation is not. This suggests that the techno-determinism of the TUA is prescriptive, not descriptive (see Section 4.3 for further discussion).

83 Ord, Mercer and Dannreuther (2021).

84 Bostrom (2019).

85 While some may choose to read Bostrom's discussion as a merely speculative proposal for a hypothetical scenario, it exemplifies a willingness to consider, and even justify, measures that would normally be deemed unconscionable in the name of averting supposed existential catastrophes. Moreover, the fact that the essay was published in a journal for policy (with policy implications considered), rather than philosophy, suggests that the discussion should not be considered as entirely theoretical and not

actionable. Additionally, intentional outreach through a TED Talk (with over 2.5 million views), the Sam Harris podcast, and a piece in Aeon all stress this aspect of the article. These actions signal that the argument is not intended to be an innocuous thought experiment.

86 The term *existential ethics* is already sometimes used in relation to the philosophy of existentialism. We are aware of this; however, it seems difficult to move away from the use of the existential given the field's reliance on the term existential.

87 Purtill, Corinne. 'How close is humanity to the edge?', *The New Yorker* (2020). https://www.newyorker.com/culture/annals-of-inquiry/how-close-is-humanity-to-the-edge

88 We owe this point to Joshua Teperowski Monrad.

89 Beard et al. (2021).

90 They are, in Bostrom's words, "mere ripples on the surface of the great sea of life" (Bostrom, 2002).

91 Bostrom (2013).

92 Bostrom.

93 Ord (2020).

94 Silver, Nate. *The Signal and the Noise: The Art and Science of Prediction*. Penguin Economics (2013).

95 Smil, Vaclav. *Growth: From Microorganisms to Megacities*. The MIT Press (2020).

96 Scheidel, Walter. *The Great Leveller: Violence and the History of Inequality: From the Stone Age to the Twenty-First Century*. Princeton University Press (2017).

97 Halstead, John. 'Is climate change an existential risk?', *Google Doc* (blog) (2019). https://docs.google.com/document/d/1qmHh-cshTCMT8LX0Y5wSQm8FMBhaxhQ8OlOeRLkXIF0/edit#; Piper, Kelsey. 'Is climate change an "existential threat" — or just a catastrophic one?', *Vox* (2019). https://www.vox.com/future-perfect/2019/6/13/18660548/climate-change-human-civilization-existential-risk. Halstead does not provide a definition of existential risk, so it is unclear what this means.

98 Pescaroli, Gianluca and David Alexander. 'Understanding compound, interconnected, interacting, and cascading risks: A holistic framework for understanding complex risks', *Risk Analysis, 38*(11) (November 2018): 2245–57. https://doi.org/10.1111/risa.13128

99 Pescaroli and Alexander (2018); Tang, Aaron and Luke Kemp. 'A fate worse than warming? Stratospheric aerosol injection and catastrophic risk', *Frontiers in Climate, 3*(720312) (2021). https://doi.org/10.3389/fclim.2021.720312

100 Defined as being consistent with our background knowledge (Betz, 2016).

101 "The case for existential risk from AI is clearly speculative. Indeed, it is the most speculative case for a major risk in this book" (Ord, 2020, p.149).

102 Cremer (2021); Grace et al. (2018).

103 Bostrom (2016); Ord (2020).

104 Centeno, Miguel A. et al. 'The emergence of global systemic risk', *Annual Review of Sociology, 41*(1) (14 August 2015): 65–85. https://doi.org/10.1146/annurev-soc-073014-112317

105 Challinor, Andy J. et al. 'Transmission of climate risks across sectors and borders',

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2121) (13 June 2018): 20170301. https://doi.org/10.1098/rsta.2017.0301

106 Homer-Dixon, Thomas et al. 'Synchronous failure: The emerging causal architecture of global crisis', *Ecology and Society, 20*(3) (2015): art6. https://doi.org/10.5751/ES-07681-200306

107 IPCC. *The Concept of Risk in the IPCC Sixth Assessment Report: A Summary of Cross-Working Group Discussions*. Intergovermental Panel on Climate Change (2020). https://www.ipcc.ch/site/assets/uploads/2021/02/Risk-guidance-FINAL_15Feb2021.pdf

108 Simpson, Nicholas P. et al. 'A framework for complex climate change risk assessment', *One Earth, 4*(4) (April 2021): 489–501. https://doi.org/10.1016/j.oneear.2021.03.005. In disaster risk management even more risk components have been proposed, such as capacity. We highlight this formulation due to the prominence and rigour of the IPCC.

109 Mani, Lara, Asaf Tzachor and Paul Cole. 'Global catastrophic risk from lower magnitude volcanic eruptions', *Nature Communications, 12*(1) (December 2021): 4756. https://doi.org/10.1038/s41467-021-25021-8

110 The injection of aerosols into the stratosphere to reflect sunlight and mitigate some of the effects. of climate change.

111 Tang and Kemp (2021).

112 Rees (2004); Ord (2020).

113 Ord (2020).

114 Avin et al. (2018); Cotton-Barratt, Owen, Max Daniel and Anders Sandberg. 'Defence in depth against human extinction: Prevention, response, resilience, and why they all matter', *Global Policy, 11*(3) (May 2020): 271–82. https://doi.org/10.1111/1758-5899.12786

115 Rees (2004); Ord (2020); Bostrom (2013).

116 Bostrom, Nick. 'The future of humanity', in *New Waves in Philosophy of Technology*, ed. Jan Kyrre Berg Olsen, Evan Selinger and Søren Riis. Palgrave Macmillan (2009), pp.186–215. https://doi.org/10.1057/9780230227279_10

117 Dafoe, Allan. 'On technological determinism: A typology, scope conditions, and a mechanism', *Science, Technology, & Human Values, 40*(6) (November 2015): 1047–76. https://doi.org/10.1177/0162243915579283

118 Balwit, Avital. 'Response to recent criticisms of longtermism', *Effective Alturism Forum* (blog) (2021).

119 Ord (2020).

120 Dafoe (2015).

121 Jacobsen, Annie. *The Pentagon's Brain: An Uncensored History of DARPA, America's Top Secret Military Research Agency* (1st ed.). Little, Brown and Company (2015).

122 Fleming, James Rodger. 'The pathological history of weather and climate modification: Three cycles of promise and hype', *Historical Studies in the Physical and Biological Sciences, 37*(1) (1 September 2006): 3–25. https://doi.org/10.1525/hsps.2006.37.1.3

123 Mumford, Lewis. *Technics and Civilization*. The University of Chicago Press (2010).

124 Malm, Andreas. *Fossil Capital: The Rise of Steam-Power and the Roots of Global Warming*. Verso (2016).

125 Bostrom (2002).

126 Bostrom (2013).

127 Millett, Piers and Andrew Snyder-Beattie. 'Existential risk and cost-effective biosecurity', *Health Security, 15*(4) (August 2017): 373–83. https://doi.org/10.1089/hs.2017.0028

128 Singh, Riddhi, Patrick M. Reed and Klaus Keller. 'Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response', *Ecology and Society, 20*(3) (2015): art12. https://doi.org/10.5751/ES-07687-200312

129 Ord (2020).

130 Morgensen, Andreas. 'Maximal cluelessness', *GPI Working Paper, 2*. University of Cambridge, Global Priorities Institute (2019); Morgensen, Andreas and David Thorstad. 'Tough enough? Robust satisficing as a decision norm for long-term policy analysis', *GPI Working Paper, 15*. University of Cambridge, Global Priorities Institute (2020).

131 Kokotajlo, Daniel. 'Tiny probabilities of vast utilities: Bibliography and appendix', *Effective Alturism Forum* (blog) (2018). https://forum.effectivealtruism.org/s/ji4eCGYmppjjWY6yp/p/nMfhXPiqPRDtEdauY#Academic_literature_; Bostrom, Nick. 'Pascal's mugging', *Analysis, 69*(3) (2009): 443–45.

132 Wilkinson, Hayden. 'In defense of fanaticism', *Ethics, 132*(2) (1 January 2022): 445–77. https://doi.org/10.1086/716869

133 Bostrom (2009).

134 Balfour, Dylan. 'Pascal's mugger strikes again', *Utilitas, 33*(1) (March 2021): 118–24. https://doi.org/10.1017/S0953820820000357

135 Based on personal communications with several scholars in the field.

136 Wilkinson (2022).

137 Betz, Gregor. 'Accounting for possibilities in decision making', in *The Argumentative Turn in Policy Analysis* (vol. 10), ed. Sven Ove Hansson and Gertrude Hirsch Hadorn. Springer International Publishing (2016), pp.135–69. https://doi.org/10.1007/978-3-319-30549-3_6

138 IPCC (2020); Simpson et al. (2021).

139 Snyder-Beattie, Andrew E., Toby Ord and Michael B. Bonsall. 'An upper bound for the background rate of human extinction', *Scientific Reports, 9*(1) (December 2019): 11054. https://doi.org/10.1038/s41598-019-47540-7

140 Ord (2020). These figures are not water tight: they assume that the vulnerability of humanity in the Palaeolithic is identical to today.

141 Ord (2020).

142 Bostrom (2002).

143 Rees (2004).

144 Ord (2020).

145 Bostrom (2019).

146 Ord (2020); Bostrom (2002).

147 That is, a superintelligent AI system that is aligned with human values and will aid rather than harm humanity.

148 Bostrom (2002).

149 Deudney, Daniel. *Dark Skies: Space Expansionism, Planetary Geopolitics, and The Ends of Humanity*. Oxford University Press (2020).

150 Harari, Yuval Noaḥ. *Homo Deus: A Brief History of Tomorrow* (revised edition). Vintage (2017); McKibben, Bill. *Enough: Staying Human in an Engineered Age*. Owl Books (2004).

151 Kemp, Luke et al. 'Bioengineering horizon scan 2020', *Elife, 9* (2020): e54489. https://doi.org/10.7554/eLife.54489; Wintle, Bonnie C. et al. 'A transatlantic perspective on 20 emerging issues in biological engineering', *Elife, 6* (14 November 2017): e30247. https://doi.org/10.7554/eLife.30247

152 Wilkinson, Richard G. and Kate Pickett. *The Spirit Level: Why Equality Is Better for Everyone [with a New Chapter Responding to Their Critics]*. Penguin Books (2010); Wilkinson, Richard G. and Kate Pickett. *The Inner Level: How More Equal Societies Reduce Stress, Restore Sanity and Improve Everyone's Well-Being*. Allen Lane (2018).

153 Kemp, Luke et al. *Emerging Technologies and Dual-Use Concerns: A Horizon Scan for Global Public Health*. World Health Organization (2021).

154 Humphreys, S. 'Legalizing lawlessness: On Giorgio Agamben's state of exception', *European Journal of International Law, 17*(3) (1 June 2006): 677–87. https://doi.org/10.1093/ejil/chl020; Keith, Linda Camp and Steven C. Poe. 'Are constitutional state of emergency clauses effective? An empirical exploration', *Hum. Rts. Q*. 26 (2004): 1071.

155 Ferejohn, John and Pasquale Pasquino. 'The law of the exception: A typology of emergency powers', *International Journal of Constitutional Law, 2*(2) (2004): 210–39.

156 De Wilde, Marc. 'Just trust us: A short history of emergency powers and constitutional change', *Comparative Legal History, 3*(1) (2 January 2015): 110–30. https://doi.org/10.1080/2049677X.2015.1041728

157 Clarke, Lee. 'Panic: Myth or reality?', *Contexts, 1*(3) (August 2002): 21–26. https://doi.org/10.1525/ctx.2002.1.3.21; Clarke, L. and C. Chess, 'Elites and panic: More to fear than fear itself', *Social Forces, 87*(2) (1 December 2008): 993–1014. https://doi.org/10.1353/sof.0.0155; Bregman, Rutger. *Humankind: A Hopeful History*. Bloomsbury Publishing (2021); Solnit, Rebecca. *A Paradise Built in Hell: The Extraordinary Communities That Arise in Disasters*. Penguin Random House (2009).

158 Bjjrnskov, Christian and Stefan Voigt. 'More power to government = more people killed? On some unexpected effects of constitutional emergency provisions during natural disasters', *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3189749

159 Kemp, Luke. 'The "stomp reflex": When governments abuse emergency powers', *BBC Future* (2021). https://www.bbc.com/future/article/20210427-the-stomp-reflex-when-governments-abuse-emergency-powers

160 Scarry, Elaine. *Thermonuclear Monarchy: Choosing Between Democracy and Doom* (1st ed.). W. W. Norton & Company (2014).

161 Wellerstein, Alex. *Restricted Data: The History of Nuclear Secrecy in the United States*. The University of Chicago Press (2021).

162 It is worth noting that both neorealist assumptions about international relations and game theoretic approaches appear frequently in research under the TUA.

163 Singer, Peter. 'The hinge of history', *Project Syndicate* (2021). https://www.project-syndicate.org/commentary/ethical-implications-of-focusing-on-extinction-risk-by-peter-singer-2021-10.; Torres, Phil. 'Against longtermism', *Aeon* (2021). https://aeon.co/essays/why-longtermism-is-the-worlds-most-dangerous-secular-credo

164 Ord (2020).

165 Seger, Elizabeth et al. *Tackling Threats to Informed Decision-Making in Democratic Societies: Promoting Epistemic Security in a Technologically-Advanced World*. Apollo — University of Cambridge Repository (14 October 2020). https://doi.org/10.17863/CAM.64183

166 Bostrom (2019).

167 Bostrom, Nick and Matthew Van Der Mewe. 'How vulnerable is the world?', *Aeon* (2019). https://aeon.co/essays/none-of-our-technologies-has-managed-to-destroy-humanity-yet

168 Bostrom (2019).

169 Bostrom and Van Der Mewe (2019).

170 Bostrom (2019).

171 Bostrom (2002).

172 Bostrom.

173 Granick, Jennifer Stisa. *American Spies: Modern Surveillance, Why You Should Care, and What To Do About It*. Cambridge University Press (2017); Kirchner, Lauren. 'What's the evidence mass surveillance works? Not much', *Propublica* (2015). https://www.propublica.org/article/whats-the-evidence-mass-surveillance-works-not-much

174 Kemp, Luke. 'Agents of doom: Who is creating the apocalypse and why', *BBC Future* (2021). https://www.bbc.com/future/article/20211014-agents-of-doom-who-is-hastening-the-apocalypse-and-why?ocid=twfut

175 Bostrom and Van Der Mewe (2019).

176 Ord (2020).

177 Ord.

178 Landemore (2020); Landemore, Hélène. *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*. Princeton University Press (2017).

179 Hong, L. and S. E. Page. 'Groups of diverse problem solvers can outperform groups of high-ability problem solvers', *Proceedings of the National Academy of Sciences, 101*(46) (16 November 2004): 16385–89. https://doi.org/10.1073/pnas.0403723101; Hong, Lu and Scott E. Page. 'Problem solving by heterogeneous agents', *Journal of Economic Theory, 97*(1) (March 2001): 123–63. https://doi.org/10.1006/jeth.2000.2709

180 Landemore (2017).

181 OECD. *Innovative Citizen Participation and New Democratic Institutions: Catching the Deliberative Wave* (2020). https://www.oecd-ilibrary.org/governance/innovative-citizen-participation-and-new-democratic-institutions_339306da-en

182 Burchi, Francesco. 'Democracy, institutions and famines in developing and emerging countries', *Canadian Journal of Development Studies/Revue Canadienne d'études Du Développement, 32*(1) (March 2011): 17–31. https://doi.org/10.1080/02255189.2011.576136

183 Massing, Michael. 'Does democracy avert famine?', *The New York Times* (2003). https://www.nytimes.com/2003/03/01/arts/does-democracy-avert-famine.html

184 Dafoe, Allan, John R. Oneal, and Bruce Russett. 'The democratic peace: Weighing the evidence and cautious inference', *International Studies Quarterly, 57*(1) (March 2013): 201–14. https://doi.org/10.1111/isqu.12055

185 Barnhart, Joslyn N. et al. 'The suffragist peace', *International Organization, 74*(4) (2020): 633–70. https://doi.org/10.1017/S0020818320000508

186 Kull, Steven. 'Survey says: Americans back arms control', *Arms Control Today, 34*(5) (2004): 22; Løvold, Magnus. 'Lessons from the ICRC's "Millennials on War" survey for communication and advocacy on nuclear weapons', *Journal for Peace and Nuclear Disarmament, 3*(2) (2 July 2020): 410–17. https://doi.org/10.1080/25751654.2020.1859216; The Simons Foundation, *Global Public Opinion on Nuclear Weapons*. The Simons Foundation (2007).

187 UNDP. *The G20 Peoples' Climate Vote 2021*. United Nations Development Programme and the University of Oxford Department of Sociology (2021). https://www.undp.org/publications/g20-peoples-climate-vote-2021; European Commission. *Citizen Support for Climate Action: 2021 Survey*. European Commission (2021). https://ec.europa.eu/clima/citizens/citizen-support-climate-action_en; Marlon, Jennifer et al. *Yale Climate Opinion Maps 2020*. Yale Program on Climate Change Communication (2020). https://climatecommunication.yale.edu/visualizations-data/ycom-us/

188 OECD (2020).

189 Convention Citoyenne pour le Climat. *Les Propositiones de La Convention Citoyenne Pour Le Climat* (2021). https://propositions.conventioncitoyennepourleclimat.fr/pdf/ccc-rapport-final.pdf

190 Lynas, Mark. *The God Species: How Humans Really Can Save the Planet*. Fourth Estate (2012).

3. Classifying Global Catastrophic Risks

Shahar Avin, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland and Martin J. Rees

Highlights:

- This chapter presents a novel classification framework for Global Catastrophic Risk scenarios according to critical system affected, global spread mechanism, and prevention and mitigation failure.
- Extending beyond existing work that identifies individual risk scenarios, the classification system highlights convergent risk factors that merit prioritisation, and uncovers potential knowledge gaps.
- The classification system can structure an ongoing, dynamic process of knowledge aggregation and horizon scanning.
- Its proposed methodology has policy implications for research agendas and provides an interdisciplinary structure for mapping and tracking the multitude of factors that could contribute to Global Catastrophic Risks.

This chapter reproduces a paper first published in *Futures* in 2018 that has provided a conceptual and methodological framework for much of CSER's research. There are other taxonomies and methods of assessing the general characteristics of extreme global risks, and the policies put

https://doi.org/10.11647/OBP.0360.03

in place to address them, throughout this volume — for example, in Chapter 20, which presents a cartography of GCR governance.

1. Introduction

In our uncertain times it is good to have something we can all agree on: global catastrophes are undesirable. As our science advances we gain a better understanding of a broad class of Global Catastrophic Risk (GCR) scenarios that could, in severe cases, take the lives of a significant portion of the human population, and may leave survivors at enhanced risk by undermining global resilience systems.[1] Much progress has been made in identifying individual GCR scenarios, and in compiling lists of the scenarios of greatest concern, but there is currently no known methodology for compiling a comprehensive, interdisciplinary view of severe Global Catastrophic Risks. While a fully complete list of GCRs may remain beyond reach, we present here a classification framework designed specifically to draw on as broad a knowledge base as possible, to highlight commonalities between risk scenarios and identify gaps in our collective knowledge regarding Global Catastrophic Risks.

To date, research on Global Catastrophic Risk scenarios has focused mainly on tracing a causal pathway from a catastrophic event to global catastrophic loss of life.[2] Such research has been fruitful in identifying and assessing a range of such GCR scenarios. Some severe GCR scenarios have posed a persistent threat to humanity since our emergence as *Homo sapiens* (e.g. impact by a 10 km astronomical object, or a volcanic super-eruption of 1000 km^3 of tephra). Other scenarios have increased in likelihood following human population expansion and the accompanying increase in resource demands (e.g. natural pandemics or ecosystem collapse). In addition, novel GCR scenarios can accompany new technologies: some of these are relatively well established (e.g. "nuclear winter" or an engineered pandemic); others are more speculative (e.g. accidents in or weaponisation of advanced artificial intelligence, or environmental shocks from ill-judged geoengineering efforts aimed at mitigating climate change).

However, compiling a comprehensive list of plausible GCR scenarios requires exploring the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios that are typically addressed by single disciplines. The classification framework presented here breaks down the analysis of GCR scenarios into three key components: (i) a critical system (or systems) whose safety boundaries are breached by a potential threat, (ii) the mechanisms by which this threat might spread globally and affect the majority of the human population, and (iii) the manner in which we might fail to prevent or mitigate both (i) and (ii). For example, a major astronomical impact may lead to a global catastrophe if we lack the technology to deflect it (mitigation failure), *and* it raises a cloud of dust that spreads around the world (global spread mechanism), *and* that cloud of dust blocks sunlight for a sufficient length of time to undermine the global food system in a manner that we cannot overcome (critical system affected). Other scenarios will have different combinations of one or more mitigation failures, one or more global spread mechanisms, and one or more critical system breaches.

In order to gain a holistic picture of potential global catastrophes, knowledge about each of the three system components needs to be explored and shared. By first constructing a classification from the broad range of known critical systems, global spread mechanisms, and prevention and mitigation failures, and then by classifying known GCR scenarios according to these dimensions, we aim to: (i) showcase the GCR relevance of a variety of scientific disciplines, (ii) highlight how commonalities between threat scenarios have research and policy implications, and (iii) highlight areas where there are potential gaps in our knowledge of global catastrophic risks. We also propose concrete steps for coordinating the broad-based, interdisciplinary research required to meet the challenges highlighted by the framework.

2. Critical Systems

We define a "critical system" as any system or process that, if disturbed beyond a certain limit or scale, could trigger a significant reduction in humanity's ability to survive in its current form (see Figure 1).

Building on the "life support systems" outlined in the research on so-called planetary boundaries (many of which appear in our *biogeochemical* group),[3] and their potential links to GCRs,[4] we identify critical systems and processes that, if disrupted, would affect human ability to survive. While we aim for comprehensiveness and minimal overlap, we acknowledge that different systems overlap. For example, while the processes affecting *ocean acidity* have direct effects on ecosystem stability and thus human life, there is significant overlap (causally, structurally and academically) with the global *water cycle, carbon cycle* and *sulphur cycle* systems.

In our classification framework, critical systems are grouped at different levels in a hierarchy, such that "higher-level" systems rely on the functioning of those at a "lower level". Thus, the framework builds up from the stability of life-supporting *physical* systems, through *cellular* and other systems, right up to species-wide *ecological* and *sociotechnological* systems. "Lower-level" systems are directly linked to human survival (which relies on functioning *anatomical* systems, which in turn relies on *cellular* systems, etc.). "Higher-level" systems, especially technology-enabled ones such as the *food* and *health* systems, help maintain the human population at its current size, and provide resilience. If these "higher-level" systems were to be disturbed significantly in some scenario — e.g. through a severe and prolonged disruption to utilities networks (such as water and electricity), or through shock effects (such as social unrest) — these could cause more harm than the system disturbance itself.

Identification of critical systems, and their cross-links, could also come from historical and archaeological study of more limited instances of human population collapse. For instance, the collapse of the Easter Island civilisation shows how excessive *resource extraction* (of palms for the making of canoes) led to ecological degradation, undermining *primary production* and *food chains*, which in turn led to failure of the Easter Island society's *food* system.[5] Further study of each critical system requires specialised expertise, often in more than one domain, as there is no one-to-one mapping from scientific disciplines to critical systems. Future work, conducted with collaboration with the wider scientific community, could lead to the demarcation of safe operating bounds for each critical system, following the example of Rockström et al.[6]

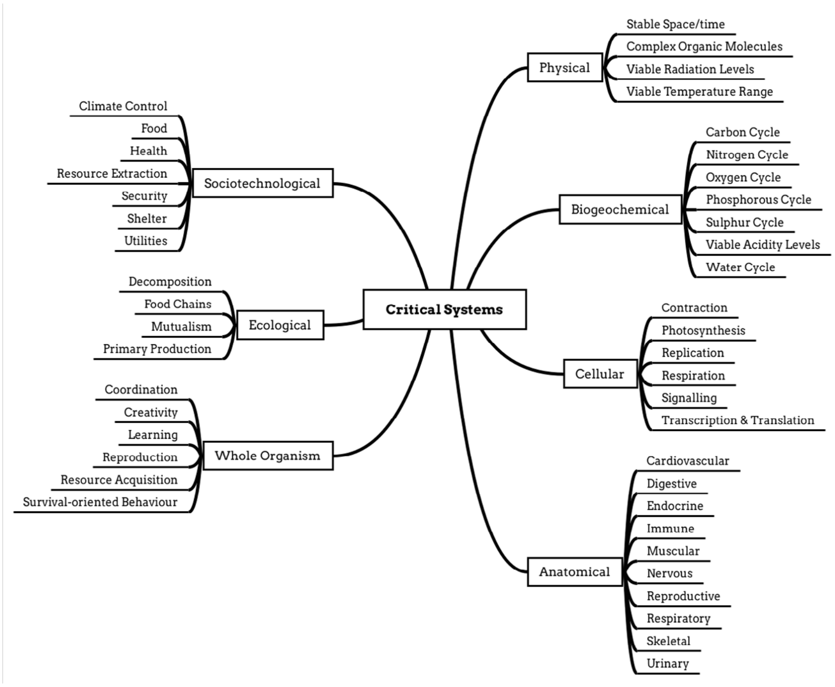


Fig. 1: Classification of critical systems aimed at identifying Global Catastrophic Risk scenarios. Systems are grouped at different levels, arranged from "lower level" to "higher level" in a clockwise fashion starting with the "physical" group on the top right.

3. Global Spread Mechanisms

For many critical systems, a failure of some instances of the system, e.g., regional crop failure, would fall far short of posing a GCR. In severe GCR scenarios, the failure of critical systems is coupled with some mechanism by which this failure spreads globally, thus potentially threatening the majority of the human population. In the framework, we separate the analysis of global spread mechanisms from the analysis of critical systems (Figure 2). This separate focus on global spread allows us to identify relevant mechanisms (and means to manage or control them) as targets of study meriting further attention, and highlights interesting commonalities.

A critical system failure can spread globally without human intervention: some *astronomical* objects or events are sufficiently massive to have direct global effect, while other threats can spread through the dynamic systems of the *natural* environment, such as the *air*- and *water-based dispersal* systems. Dust and toxins could be spread naturally even if they do not replicate, though of course a self-replicating threat (e.g. a virus that affects multiple species of fish) could couple with a dynamic system (e.g. ocean currents) to achieve much faster spread.

In addition to natural spread, many risk scenarios, and especially emergent risk scenarios, rely on the highly connected nature of our species, both materially and conceptually. A modern pandemic can spread through airports and other mass-transit hubs of the globe-encompassing *transit* network, thus coupling a *biological replicator* (this might be, e.g., a bacterium itself, *or* a biological vector, e.g. a mosquito) to a highly connected *anthropogenic network*. A cyber attack can cascade through global critical systems at the speed of *digital communication*, shutting down *health* and *security* systems, and undermining *resource extraction* and *utilities* by disrupting mines and power plants (a *digital replicator*, such as a computer worm, could speed up the spread rate and reach).

Access to information can play a more abstract, but no less important, role in the spread of critical system failure. The widespread, and growing, access of individuals and groups across the globe to ideas, schematics, and manufacturing capabilities (e.g. Do-It-Yourself, or DIY, biology) through *digital* and *cultural* exchanges (e.g. online fora), enables novel hypothetical GCR scenarios. Such a scenario could start with, say, the accidental or malicious release of a home-grown pathogen, or the one-sided deployment of geoengineering efforts in an attempt to mitigate climate change. Some ideas encourage their own spread, e.g. schematics for communication devices, or ideas that encourage further sharing of those ideas (e.g. ideologies or viral videos), coupling *cultural replicators* with human interaction networks.

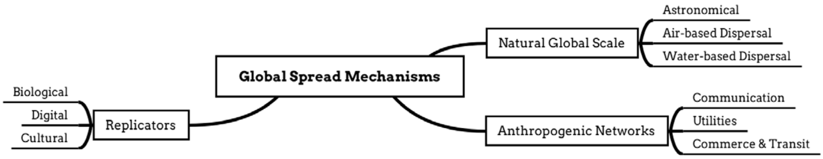
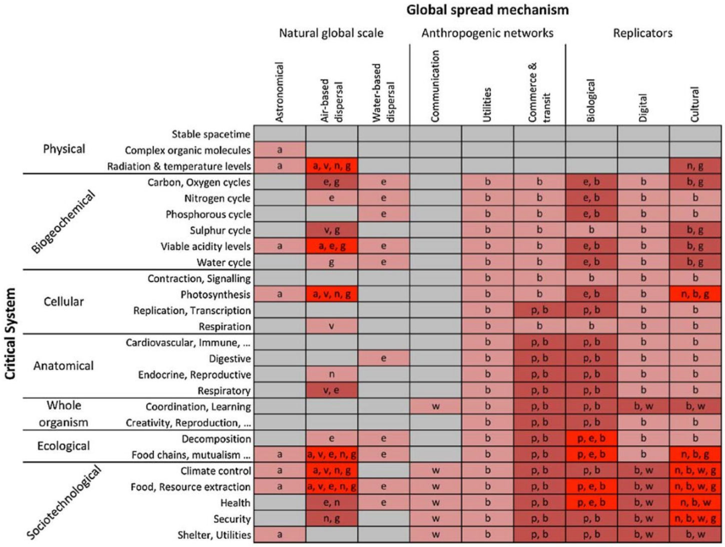


Fig. 2: Classification of global spread mechanisms relevant to Global Catastrophic Risk.

Table 1. Classification of hypothetical global catastrophic risk scenarios by global spread mechanisms and critical systems affected. Letters represent eight examples of risk scenarios: asteroid impact (a), volcanic super-eruption (v), pandemic (natural) (p), ecosystem collapse (e), nuclear war (n), bioengineered pathogen (b), weaponised artificial intelligence (w), geoengineering termination shock (g). Cell colour represents number of catastrophic scenarios potentially compromising the critical system globally via the spread mechanism (grey: no likely disruption, light pink: one scenario, dark pink: two scenarios, red: three or more scenarios). Critical systems with an identical vulnerability profile to these risk scenarios have been omitted for brevity, indicated by ellipses (see Fig. 1 for the full list of systems). (For interpretation of the references to colour in this table legend, the reader is referred to the web version of this article.)



		Global spread mechanism								
		Natural global scale			Anthropogenic networks			Replicators		
Critical System		Astronomical	Air-based dispersal	Water-based dispersal	Communication	Utilities	Commerce & transit	Biological	Digital	Cultural
Physical	Stable spacetime									
	Complex organic molecules	a								
	Radiation & temperature levels	a	a, v, n, g							n, g
Biogeochemical	Carbon, Oxygen cycles		e, g	e		b	b	e, b	b	b, g
	Nitrogen cycle		e	e		b	b	e, b	b	b
	Phosphorous cycle			e		b	b	e, b	b	b
	Sulphur cycle		v, g			b	b	b	b	b, g
	Viable acidity levels	a	a, e, g	e		b	b	e, b	b	b, g
	Water cycle		g	e		b	b	e, b	b	b, g
Cellular	Contraction, Signalling					b	b	b	b	b
	Photosynthesis	a	a, v, n, g			b	b	e, b	b	n, b, g
	Replication, Transcription					b	p, b	p, b	b	b
	Respiration		v			b	b	b	b	b
Anatomical	Cardiovascular, Immune, ...					b	p, b	p, b	b	b
	Digestive			e		b	p, b	p, b	b	b
	Endocrine, Reproductive		n			b	p, b	p, b	b	b
	Respiratory		v, e			b	p, b	p, b	b	b
Whole organism	Coordination, Learning				w	b	p, b	p, b	b, w	b, w
	Creativity, Reproduction, ...					b	p, b	p, b	b	b
Ecological	Decomposition		e	e		b	p, b	p, e, b	b	b
	Food chains, mutualism ...	a	a, v, e, n, g	e		b	p, b	p, e, b	b	n, b, g
Sociotechnological	Climate control	a	a, v, n, g		w	b	p, b	p, b	b, w	n, b, w, g
	Food, Resource extraction	a	a, v, e, n, g	e	w	b	p, b	p, e, b	b, w	n, b, w, g
	Health		e, n	e	w	b	p, b	p, e, b	b, w	n, b, w
	Security		n, g		w	b	p, b	p, b	b, w	n, b, w, g
	Shelter, Utilities	a			w	b	p, b	p, b	b, w	b, w

Table 1 illustrates how analysis of critical systems and analysis of global spread mechanisms might be combined into a single classification framework. The table presents a mapping from eight hypothetical GCR scenarios to the critical systems that are most likely to be undermined in each scenario, for each type of global spread mechanism. We have chosen a selection of severe GCR scenarios that are (i) familiar, (ii) considered plausible, and (iii) cover both natural and anthropogenic threats. This is far from a comprehensive list of scenarios, as the very framework presented here aims to help explore possible scenarios.

4. Prevention and Mitigation Failures

Analysing GCR scenarios along the dimensions of critical systems and spread mechanisms draws significantly on our understanding of the natural world and technical systems, and complements existing endeavours to classify risks of a smaller scale.[7] Holistic risk management, however, must take into account the human elements that moderate GCR through prevention and mitigation efforts, and how these efforts might fail. The challenge of preventing global catastrophes thus requires integration of the work and expertise in and between the natural and the social sciences, on a global scale.

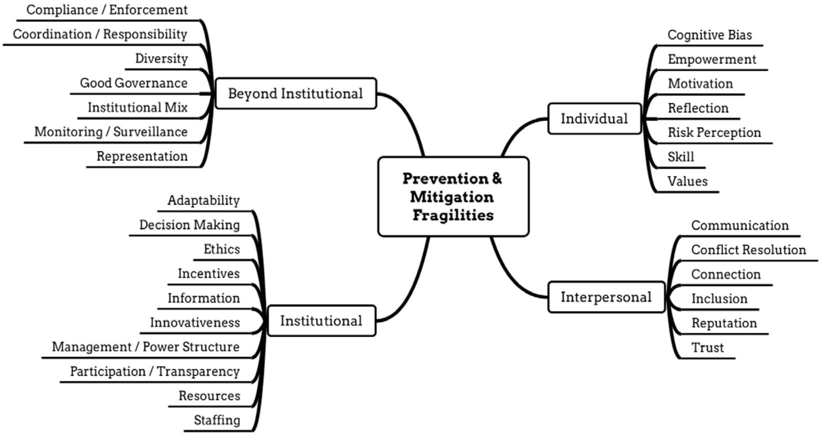


Fig. 3: Levels and dimensions of prevention and mitigation factors moderating global catastrophic risks.

A particularly comprehensive existing risk management framework with such integrative characteristics and international scope is the Sendai Framework for Disaster Risk Reduction (SFDRR), adopted by 187 UN member states in 2015.[8] Although developed for natural rather than technological disasters, it considers many of the potential human factors that influence resilience and vulnerability to an unfolding disaster. We take a similar approach here, and identify potentially fragile areas in the global risk prevention and mitigation system (Figure 3). Rather than aiming for comprehensiveness or exclusivity, it highlights that understanding these interdependent and complex human factors requires input from a wide range of disciplines beyond the natural sciences.

For instance, short-term thinking and a limited focus constitute *cognitive biases* affecting *risk perception* and *management* on the individual and institutional level (as studied in psychology and behavioural economics); *unresolved political conflicts* and competing *ethical* notions of justice undermine international *cooperation* and burden-sharing on the institutional and supra-institutional level (as studied in e.g. law, philosophy and political science).

Some risks (e.g. natural pandemics) are already the focus of well-developed institutional systems (e.g. the World Health Organization), robust research activity and technical know-how. For GCRs from emerging technologies, however, the *institutional mix* and a research agenda are only just becoming established. Conventional disaster response (e.g. recovery and compensation), and even newer, comprehensive strategies (e.g. the "build back better" principle adopted in some countries post-disaster) are inadequate for addressing threat scenarios where there is limited reaction time and no second chance. For these cases, we need a novel framework that is at least as interdisciplinary as the SFDRR, but moves away from uni-dimensional, natural hazards and instead addresses complex, anthropogenic risks, which are far more likely to cause a severe global catastrophe.[9] In particular, we have to focus on the prevention and mitigation of multidimensional risk scenarios that involve cascades of socio-technological, natural-technological ("natech") and technological-natural disasters.

As we confront emergent technological GCR scenarios, lessons can be learnt from previous smaller disasters. An instructive recent case of a multi-dimensional disaster scenario, albeit of local scope, is the Fukushima Dai'ichi nuclear accident, which laid bare failures at the interface of natural, scientific, technological, socioeconomic, legal and political realms. One such failure was the supervision of Japan's nuclear industry by the very same authorities that were to promote nuclear technology. Such an institutional setup, aggravated by *cognitive biases* (e.g. groupthink) in a sector with revolving doors to the regulator, was lacking adequate *incentive* structures, and was destined to result in conflicts of interest and regulatory capture. The international science and policy community therefore has the opportunity and the responsibility to co-create better risk prevention and mitigation systems, by engaging with researchers in the social sciences and humanities.

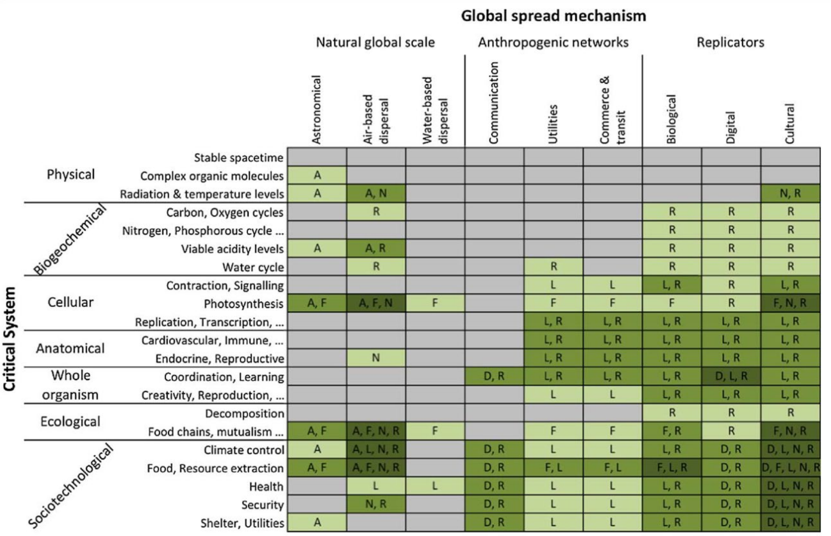
In principle it is possible to create a table that would expand on Table 1 to include the third dimension described here, i.e., prevention and mitigation failures. Such a table is, however, difficult to produce in practice, as the scenarios it helps us distinguish between are more fine-grained than those classified in Table 1. They are subcategories of these scenarios. For example, in Table 1 we classified "natural pandemic" as a single scenario, yet from a disaster policy and risk reduction perspective there is a clear difference between a pandemic that emerged due to underinvestment in veterinary surveillance, and a pandemic that emerged due to accidental release from a research laboratory. These scenarios can be further subdivided through the precise failures that allow the pandemic risk to materialise. If we consider just the accidental release scenario, we would start from the grid items occupied by "p" in Table 1, which highlight intersections of the critical systems undermined by pandemic, such as anatomical systems, and the spread mechanisms for pandemic, which naturally include biological replicators but are also affected by anthropogenic networks as well as air- and water-based dispersal. To these we would add a third dimension, that would highlight all the prevention and mitigation failures potentially involved in accidental release, from failures of individual *skill* or *risk perception*, through institutional failures including malformed *incentives*, or insufficient *staffing* and *resources*, to supra-institutional failures of insufficient *monitoring* and *enforcement*.

5. Intended Use of the Classification System

In this section, we illustrate three key ways the classification system could potentially be used, although more may be discovered as the system is expanded and updated.

The first potential use is to prioritise risk reduction efforts. As can be seen in Table 1, scenarios with significantly different primary causes could manifest their GCR potential through a similar mechanism. For example, asteroid impact, volcanic super-eruption and nuclear war scenarios all feature a risk of significant reduction of inbound solar radiation, disrupting food security and potentially leading to mass starvation. Not only does this draw attention to systems that are vulnerable to multiple hazards, but it also suggests there is value in considering these scenarios together in research and policy contexts, rather than thinking about them in isolation. For example, if accounting for volcanic super-eruptions, asteroid impacts and nuclear wars together, one might seriously consider risk management strategies that are robust to all scenarios, such as alternative food production systems to withstand the multi-year "winter" that might follow.[10] While this does not preclude investment in nuclear disarmament or asteroid deflection, it demonstrates that alternative food policies may warrant more attention than first thought.

Table 2: Classification of risk reduction strategies by global spread mechanisms and critical systems affected. Letters represent six examples of risk reduction strategies: asteroid deflection (A), digital resilience (D), food production through non-photosynthetic processes (F), limiting human contact during a pandemic (L), nuclear disarmament (N), restrictions on the diffusion of risky technologies (R). Cell colour represents number of risk reduction strategies addressing possible critical system failure and its global spread via the mechanism (grey: not addressed, light green: one strategy, green: two strategies, dark green: three or more strategies). Critical systems with an identical benefit profile from these strategies have been omitted for brevity, indicated by ellipses (see Figure 1 for the full list of systems).



			Global spread mechanism								
			Natural global scale			Anthropogenic networks			Replicators		
			Astronomical	Air-based dispersal	Water-based dispersal	Communication	Utilities	Commerce & transit	Biological	Digital	Cultural
Critical System	Physical	Stable spacetime									
		Complex organic molecules	A								
		Radiation & temperature levels	A	A, N							N, R
	Biogeochemical	Carbon, Oxygen cycles		R					R	R	R
		Nitrogen, Phosphorous cycle ...							R	R	R
		Viable acidity levels	A	A, R					R	R	R
		Water cycle		R			R		R	R	R
	Cellular	Contraction, Signalling					L	L	L, R	R	L, R
		Photosynthesis	A, F	A, F, N	F		F	F	F	R	F, N, R
		Replication, Transcription, ...					L, R	L, R	L, R	L, R	L, R
	Anatomical	Cardiovascular, Immune, ...					L, R	L, R	L, R	L, R	L, R
		Endocrine, Reproductive		N			L, R	L, R	L, R	L, R	L, R
	Whole organism	Coordination, Learning				D, R	L, R	L, R	L, R	D, L, R	L, R
		Creativity, Reproduction, ...					L	L	L, R	L, R	L, R
	Ecological	Decomposition							R	R	R
		Food chains, mutualism ...	A, F	A, F, N, R	F		F	F	F, R	R	F, N, R
	Sociotechnological	Climate control	A	A, L, N, R		D, R	L	L	L, R	D, R	D, L, N, R
		Food, Resource extraction	A, F	A, F, N, R		D, R	F, L	F, L	F, L, R	D, R	D, F, L, N, R
		Health		L	L	D, R	L	L	L, R	D, R	D, L, N, R
		Security		N, R		D, R	L	L	L, R	D, R	D, L, N, R
		Shelter, Utilities	A			D, R	L	L	L, R	D, R	D, L, N, R

In addition to the challenge of securing food under reduced solar radiation, the classification framework highlights other areas that warrant further attention as potentially occurring from a range of threats. These include: how to manage the proliferation of potentially dangerous technologies, how we would function if human contact was restricted during a pandemic spread,[11] and how we might make critical digital systems resilient to disruption by error or malice. The value of the classification system in highlighting potentially compatible risk reduction strategies is visualised in Table 2.

While expansion of this table into the third dimension of prevention and mitigation failures is beyond the scope of the current chapter, we foresee that the creation of such an expansion, in a dynamic and collaborative fashion as described below, will have the same benefits as Table 1. That is, it could be used to focus attention on prevention and mitigation failure categories that affect a range of GCR scenarios (e.g. better risk communication tools). While policy relevance to multiple risks does not directly entail higher priority for an intervention (as matters of probability, effectiveness and cost need to be taken into account), it could indicate the value of a comprehensive cross-risk analysis, to paint a more complete picture of the value of a proposed intervention.

The second potential use for the classification system lies in creating a live reference list of expertise for different risk scenarios. Our attempt to carve out categories in each dimension based on different academic domains should provide a quick index of the academic disciplines that are essential to "have at the table" when researching a specific risk scenario. Such an index could prove useful for policy-makers who take responsibility for certain risk domains, or when an emerging risk is unfolding and an interdisciplinary team needs to be assembled in a hurry. This potential use underscores the importance of including the third dimension, which points to relevant academic disaster management expertise outside the natural sciences.

The third potential use for the system is as a tool to highlight highly uncertain or neglected corners of the GCR possibility space, and guide research efforts towards these corners, in the hope of discovering unknown unknowns. The combinatorial nature of the classification systems provides a natural way of progressing from well-known systems and mechanisms to a vast and as-yet largely unexplored space of possible GCRs. Admittedly, even an exhaustive exploration of all possible GCR scenario configurations within the current classification system would not provide a guarantee against "black swans", but it can certainly foster a fuller understanding of the threats we face.

6. Where to Next?

The classification framework presented above is dynamic, spanning a broad range of disciplines and reflecting a dense web of interacting variables along three dimensions: where critical systems are vulnerable to GCRs, how threats might spread globally, and how attempts to prevent or mitigate these threats might fail due to human factors. To successfully maintain awareness and organise the plethora of knowledge around GCRs we need to meet the following challenges:

1. collect, aggregate and digest information from highly distributed knowledge networks, overcoming communication barriers and delays;
2. update regularly the classification of GCR scenarios as knowledge advances, and as technology shapes — or is poised to shape — the relevant domains.

Meeting these challenges requires a combination of strategies. It would be sensible to populate a classification framework using a group elicitation approach, calling on experts in different critical systems, global reach mechanisms and mitigation approaches to produce short summaries containing signposts to evidence in their fields that would be relevant to GCRs. Such summaries would then be aggregated in a central repository. A group of multi-domain experts could serve as editors to make sure efforts are coordinated, language is harmonised and appropriate for an interdisciplinary audience, and credit is attributed appropriately. Similar, successful repositories for other disciplines already exist and could provide inspiration.[12] The evolving classification system, when part of a knowledge synthesis effort, could offer a visual way to communicate the current state of knowledge.[13]

As the frontiers of knowledge and innovation expand, so too does the horizon of our possible futures. The framework outlined here could both inform, and be informed by, different "foresight" tools.[14] It may be a useful tool for generating scenarios that help us explore and prepare for new risks, emerging trends and key uncertainties. Scenarios can then be characterised in more detail and monitored using horizon scanning,[15] another tool in the ″foresight″ suite. Structured horizon-scanning methods could be useful to scan for the early signals of a scenario unfolding, or simply to update the classification framework with information on new discoveries, innovation, theories and data produced by the scientific community.

Globalisation and technology are advancing at a rapid pace, and it is difficult to appraise the ever-changing landscape of risks. In order for research into new, potentially disruptive technologies to proceed responsibly, and to better anticipate how interacting threats may unfold across our globe, the state of knowledge around risks and potential risk mitigation measures needs to be transparent, organised and updateable. We hope that the classification framework outlined in this chapter will facilitate the communication between disciplines that such an endeavour needs.

Notes and References

[1] Rees, M. J. *Our Final Hour*. Basic Books (2003); Posner, R. A. *Catastrophe: Risk and Response*. Oxford University Press (2004); Bostrom, N. and M. M. Ćirković. (eds.) *Global Catastrophic Risk*. Oxford University Press (2008); Tonn, B. E. and D. MacGregor. 'Human extinction' [special issue], *Futures, 41* (2009), 673–774. https://doi.org/10.1016/j.futures.2009.07.013; Baum, S. D. and B. E. Tonn. (eds.) 'Confronting future catastrophic threats to humanity' [special issue], *Futures, 72* (2015), 1–96.

[2] Asimov, I. *A Choice of Catastrophes: The Disasters That Threaten Our World*. Fawcett Columbine (1981); Bostrom, N. 'Existential risks: Analyzing human extinction scenarios and related hazards‘, *Journal of Evolution and Technology, 9* (2002); Coburn, A. W., G. Bowman, S. J. Ruffle, R. Foulser-Piggott, D. Ralph and M. Tuveson. 'A taxonomy of threats for complex risk management‘, *Cambridge Risk Framework Series*. Centre for Risk Studies, University of Cambridge (2014); Turchin, A. *How Many X-Risks for Humanity? This Roadmap Has 100 Doomsday Scenarios*. Institute for Ethics and Information Technology (2015). https://ieet.org/index.php/IEET2/more/turchin20150623; Cotton-Barratt, O., S. Farquhar, J. Halstead, S. Schubert and A. Snyder-Beattie. *Global Catastrophic Risks 2016*. Global Challenges Foundation (2016); Bostrom and Ćirković (2008).

[3] Rockström, J. et al. 'Planetary boundaries: Exploring the safe operating space for humanity', *Ecology and Society, 14*(2) (2009), p.32 https://doi.org/10.5751/ES-03180-140232; Steffen, W. et al. 'Planetary boundaries: Guiding human development on a changing planet', *Science, 347*(6223) (2015), 1259855. https://doi.org/10.1126/science.1259855

[4] Baum, S. D. and I. C. Handoh. 'Integrating the planetary boundaries and global catastrophic risk paradigms', *Ecological Economics, 107* (2014), 13–21. https://doi.org/10.1016/j.ecolecon.2014.07.024

[5] Morrison, K. D. 'Failure and how to avoid it', *Nature, 440* (2006), 752–54. https://doi.org/10.1038/440752a

[6] Rockström et al. (2009).

[7] IRDR. *Peril Classification and Hazard Glossary*. UCL Institute for Risk and Disaster Reduction (2014).

[8] UNISDR. *Sendai Framework for Disaster Risk Reduction 2015–2030*. United Nations Office for Disaster Risk Reduction (2015).

[9] Rees (2003).

[10] Denkenberger, D. and J. M. Pearce. *Feeding Everyone No Matter What*. Elsevier (2015). https://doi.org/10.1016/c2015-0-04027-8

[11] Please note that this chapter was written several years before the COVID-19 pandemic and has not been updated since. CSER is currently undertaking a multi-year project to learn lessons from the pandemic for the long term.

[12] Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. Stanford University (2016). https://plato.stanford.edu/; Wolfrum, R. (ed.) *The Max Planck Encyclopedia of Public International Law*. Oxford University Press (2017). http://opil.ouplaw.com/home/EPIL

[13] McKinnon, M. C., S. H. Cheng, R. Garside, Y. J. Masuda and D. C. Miller. 'Sustainability: Map the evidence', *Nature, 528*(7581) (2015), 185–87. https://doi.org/10.1038/528185a

[14] Cook, C. N., S. Inayatullah, M. A. Burgman, W. J. Sutherland and B. A. Wintle. 'Strategic foresight: How planning for the unpredictable can improve environmental decision-making', *Trends in Ecology & Evolution, 29* (2014), 531–41. https://doi.org/10.1016/j.tree.2014.07.005

[15] Sutherland, W. J. and H. J. Woodroof. 'The need for environmental horizon scanning', *Trends in Ecology & Evolution, 24* (2009), 523–27. https://doi.org/10.1016/j.tree.2009.04.008 van Rij, V. 'Joint horizon scanning: Identifying common strategic choices and questions for knowledge', *Science and Public Policy, 37* (2010), 7–18. https://doi.org/10.3152/030234210x484801; Amanatidou, E. et al. 'On concepts and methods in horizon scanning', *Science and Public Policy, 39* (2012), 208–21. https://doi.org/10.1093/scipol/scs017

4. Governing Boring Apocalypses: A New Typology of Existential Vulnerabilities and Exposures for Existential Risk Research

Hin-Yan Liu, Kristian Cedervall Lauta and Matthijs Michiel Maas

Highlights:

- It is unhelpful to focus on explosive catastrophes that could directly kill all humans at the same time. Other potential paths or disaster interaction effects that converge towards that same disastrous outcome, even if only indirectly or over longer timescales, are just as deadly and potentially more likely.
- We need to consider not only hazards — the external source of peril — but also vulnerabilities, propensities or weaknesses inherent within human systems that increase the likelihood of our succumbing to pressures or challenges that threaten existential outcomes, and exposures (the number, scope and nature of the interface between the hazard and the vulnerability). When studying the collapse of complex systems, such as human civilisation, vulnerabilities and exposures may be the most significant contributors to existential risk.
- Vulnerabilities and exposures can be further categorised according to whether they are ontological (exist due to the

https://doi.org/10.11647/OBP.0360.04

nature of human beings or our location in space and time); passive (exist indirectly, through lack of action); active (exist directly because of insufficient/mis-specified action); or intentional (were created on purpose).

- Understanding these different features of existential risk, their sources and characteristics not only helps us to understand the range of risks facing humanity but also to identify a wider range of policy options for risk mitigation.
- Ignoring these factors can lead us to neglect ways in which imperfect recommendations from existential risk researchers can make things worse; for instance, by playing into cultural vulnerabilities and exposures or creating "too easy" solutions that serve to block important further work.

This chapter was presented at CSER's first Cambridge Conference on Catastrophic Risk in 2016. It brings Existential Risk Studies into conversation with the field of disaster studies and sketches a now widely adopted approach to thinking about existential risk reduction based around understanding vulnerability and exposure used in several other chapters of this volume including Chapters 12 and 20. The possibility that flawed conceptions of extreme global risk could be harmful is also explored in Chapters 2 and 14.

In recent years, the study of existential risk has explored a range of natural and man-made catastrophes, from supervolcano eruption to nuclear war, and from global pandemics to potential risks from misaligned AI. What these risks have in common is that they might cause outright human extinction, were they to occur. In this approach, such identified existential risks are frequently characterised by relatively singular origin events and concrete pathways of harm which directly jeopardise the survival of humanity, or undercut its potential for long-term technological progress. While this approach aptly identifies the most cataclysmic fates which may befall humanity, we argue that catastrophic ″existential outcomes″ may likely arise from a broader range of sources and societal vulnerabilities, and through the complex interactions of disparate social, cultural and natural processes — many of which, taken in isolation, might not be seen to merit attention as a Global Catastrophic Risk, let alone an existential one.

This chapter argues that an emphasis on mitigating the hazards (discrete causes) of existential risks is an unnecessarily narrow framing of the challenge facing humanity, one which risks prematurely curtailing the spectrum of policy responses considered. Instead, it argues existential risks constitute but a subset in a broader set of challenges which could directly or indirectly contribute to existential consequences for humanity. To illustrate, we introduce and examine a set of existential risks that often fall outside the scope of, or remain understudied within, the field. By focusing on vulnerability and exposure rather than existential hazards, we develop a new taxonomy which captures factors contributing to these existential risks. Latent structural vulnerabilities in our technological systems and in our (institutional and cultural) societal arrangements (e.g. systemic "normal accidents"; institutional absence or failure; cultural distrust of authorities) may increase our susceptibility — the likelihood that we succumb to existential hazards. Finally, different types of exposure of our society or its natural base determine if or how a given hazard can interface with pre-existing vulnerabilities, to trigger emergent existential risks. We argue that far from being peripheral footnotes to their more direct and immediately terminal counterparts, these "boring apocalypses" may well prove to be the more endemic and problematic, dragging down and undercutting short-term successes in mitigating more spectacular risks. If the cardinal concern is humanity's continued survival and prosperity, then focusing academic and public advocacy efforts on reducing direct existential hazards may have the paradoxical potential of exacerbating humanity's indirect susceptibility to such outcomes.

Adopting law and policy perspectives allow us to foreground societal dimensions that complement and reinforce the discourse on existential risks. This holistic taxonomy accordingly enables scholars in the field of existential risk to better recognise the expanded range of existential risks, and helps them to better understand and deploy a more diverse toolbox of law and governance approaches to address these challenges.

Introduction: The Definition and Framings of Existential Risk

In recent years, a growing body of scholarship has argued that a new class of risks bears closer study, for their potential extreme impact on the survival of humanity.[1] Prior research has identified a range of such human extinction risks,[2] both natural and manmade, including risks from supervolcano eruption, asteroid impact, global warming, nuclear war, as well as more speculative risks from emerging technologies such as biotechnology, high-energy physics experiment disasters, or misaligned Artificial Intelligence.[3]

While it is encouraging to see greater attention for a critical topic that has long remained understudied, it is relevant to ask how the framing of the field's basic concepts shapes both the problems it identifies and prioritises, as well as the policy approaches it considers and engages. In his seminal paper, Bostrom defined an existential risk as "[o]ne where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential".[4] Thus, in Bostrom's view, existential risks are characterised both by their scope (pan-generational) and their intensity (crushing): the size of the group of people who are at risk[5] and how badly each individual within that group is affected, respectively.[6]

Much prior research on existential risks has thus deployed criteria and methodology which have identified discrete and independent challenges of sufficient severity and pervasiveness to bring about the "adverse outcome" in a direct causal manner. In this reading, existential risks are an extreme offshoot of Global Catastrophic Risks — disasters which "might have the potential to inflict serious damage to human well-being on a global scale,"[7] but which fall short of permanent collapse. While we are not necessarily averse to the Bostromian definition of "adverse outcomes" — a definition which indeed seems to characterise the space of eventual outcomes to be avoided — we take more issue with the limited range of pathways towards this dreaded outcome-space, which much of the literature has focused on exploring. Specifically, as noted by others in the community, much prior research "has focused mainly on tracing a causal pathway from a catastrophic event to global catastrophic loss of life".[8] As such, there remains an event-focus, in the sense that only discrete events that are causally connected to the demise of humanity within a relatively short time-frame qualify as an existential risk (rather than a "merely" globally catastrophic one, or a background risk).

Existential Risk (Re)Framings as Crucial Consideration for Law and Governance Approaches

Distinguishing existential risks as a uniquely threatening outlier along the spectrum of global risks, however, is arguably an unnecessarily narrow framing of the field of study. Indeed, a high-profile "one-hit-KO" existential risk such as a global nuclear war or a pandemic may constitute only one avenue towards that "adverse outcome", and concentrating predominately upon (ways to intervene in) its origin and direct pathway risks overshadowing other potential paths or disaster interaction effects[9] that functionally converge towards that same disastrous outcome, even if only indirectly or over longer timescales, with a potentially higher probability. Indeed, as recently noted by scholars in the field, a full mapping of scenarios that lead to catastrophic outcomes "requires exploring the interplay between many interacting critical systems and threats, beyond the narrow study of individual scenarios that are typically addressed by single disciplines".[10] The precise framing of "existential risks" is therefore a crucial consideration, informing ethical, strategic, and epistemological (cf. academic) priorities in facing "adverse outcomes". This is particularly the case in the context of studying how global political dynamics may interact with certain existential risks, and in formulating meaningfully effective policies and governance approaches to such risks. Of course, this is not to say that the field of existential risk studies has not sought to involve and engage with policy and governance approaches and solutions. Indeed, to its credit, research in the field of existential risks has actively sought to engage with these issues — given that, as Bostrom himself observes,[11] global cooperation is critical to mitigating a wide range of existential risks. Likewise, researchers within the "AI safety" community are beginning to highlight fields such as policy and psychology as under-represented but potentially promising approaches to addressing risks arising from AI.[12]

Accordingly, there has been research into the interaction effects between technologies and politics — such as the possibility that arms races might increase the risk that untested, powerful AI systems are deployed rashly or prematurely.[13] Other work has drawn on cognitive psychology to study how people might structurally (mis)judge the probability of risks.[14] Likewise, work exploring policy and governance approaches to mitigating existential risks has looked at insurance arrangements for large catastrophes,[15] technology taxes and subsidies,[16] and work drawing on social (and organisational) psychology to assess ways to motivate AI researchers to choose beneficial AI designs.[17] Yet, other work has examined the cost-effectiveness of biosecurity interventions;[18] pricing externalities to balance public the risks and benefits of scientific research generally;[19] and proposing a general international regulatory regime to govern global catastrophic and existential risks from emerging technologies.[20] At present, a majority of existential risk research centres[21] have articulated law and policy research as areas of interest, and scholars in this space have begun to translate such work into concrete proposed policy interventions — notably the 2017 GPP report, which included proposals to develop governance for geoengineering research, establish international scenario plans and exercises for engineered pandemics, and build international attention for existential risk reduction.[22]

Such work is highly encouraging, and the existential risk research agenda has benefited from it. Nonetheless, the risk remains that a too-narrow conception of "existential risks" prematurely closes down the space of law and governance solutions that are possible — or necessary — in assuring humanity a non-catastrophic future — for instance, a future that, in Bostrom's framing, meaningfully "maximize[s] the probability of an ok outcome".[23] However, if human extinction and the persistent and pervasive truncation of technological potential are not completely homologous, then tailoring our portfolio of policy responses exclusively to closing off the pathways these risks could take — and then calling it a day — would be insufficient. In fact, this might only afford future policy-makers with a false sense of security, even as the world continues to reside in an overall state of "super-risk".[24]

This is especially the case when there is a narrow "technological" (re)solution on offer — such as "improve global vaccine synthesis and production capability", or "subsidize international technical AI safety research" — which promise to address or prevent the risk at its root. While such direct technological solutions may certainly be indispensable to averting some existential risks, they may not suffice in actually "plugging all the holes" in our risk space. In a disciplinary context, there is a risk (admittedly self-correcting, given publication incentives) of the research agenda "halting" early. In a real-world context, the availability of simple, straightforward "fixes" might even pose a "moral hazard", if policy-makers or global governance systems which lack political will or the attention to explore more complex or costly changes seize upon the "symbolic action" of the straightforward, first-order mitigation strategies. Even where this is not the case, certain policy recommendations to mitigate existential risks might depend on too optimistic a view of institutional rationality or capability.

"Boring Apocalypses": From Existential Hazards to Existential Risks

While such efforts might mitigate specific existential risks, this might not translate into significantly lowering the overall probability of the "adverse outcome", if only a part of the problem, or only one problem among many, is addressed. An alternative articulation is that only one path to the "adverse outcome" is being explored by much research into existential risks: erecting obstacles along that path may indeed reduce the overall likelihood of manifesting these risks, but this might have little impact, or even no effect, upon the manifestation of the "adverse outcome".

Thus, our view is that a materialised existential risk (what we call an "existential hazard") is *sufficient* to lead to an (existentially) "adverse outcome", but crucially, that this is *unnecessary* to reach that result. If the overarching objective is to lower the probability of human extinction or significant technological curtailment, adopting an array of approaches which complement the mitigation of direct existential risks is required. Within this broad spectrum of aligned approaches, we propose to introduce law, policy, regulatory and governance tools in this chapter as an example. The choice of law and policy perspectives is two-fold: on one hand, they make it possible to take second-order considerations, which take indirect and socially and culturally mediated paths towards "adverse outcomes" into account; on the other hand, these recognise both the complexity of social organisation and the prospect that civilisational collapse may trigger or possibly instantiate existential outcomes. In this sense, law and policy approaches offer the possibility of complementing and enhancing the narrower approach adopted by contemporary existential risk research, to take into consideration other paths to existentially adverse outcomes; and to better anticipate vulnerabilities, exposures and failure modes in societal efforts to address existential risks.

Exploring the Implications of the Existential Risk Framing: Risks from AI

An example of this can be drawn from the prospect of superintelligent Artificial Intelligence.[25] Although the landmark research agenda articulated by Russel et al. (2015) does call for research into "short-term" policy issues, debates in this field of AI risk[26] have — with some exceptions — identified the core problem as one of value alignment, where the divergence between the interests of humanity and those of the superintelligence would lead to the demise of humanity through mere processes of optimisation. Thus, the existential risk posed by the superintelligence lies in the fact that it will be more capable than we can ever be; human beings will be outmanoeuvred in attempts at convincing, controlling or coercing that superintelligence to serve our interests. As a result of this framing, the research agenda on AI risk has put the emphasis on evaluating the technical feasibility of an "intelligence explosion"[27] through recursive self-improvement after reaching a critical threshold;[28] on formulating strategies to estimate timelines for the expected technological development of such "human-level" or "general" machine intelligence;[29] and on formulating technical proposals to guarantee that a superintelligence's goals or values will remain aligned with those of humanity — the so-called superintelligence "Control Problem".[30]

While this is worthwhile and necessary to address the potential risks of advanced AI, this framing of existential risks focuses on the most direct and causally connected existential risk posed by AI systems.

Yet while super-human intelligence might surely suffice to trigger an existential outcome, it is not necessary to it. Cynically, mere human-level intelligence appears to be more than sufficient to pose an array of existential risks.[31]

Furthermore, some applications of "narrow" AI, which might help in mitigating against some existential risks, might pose their own existential risks when combined with other technologies or trends, or might simply lower barriers against other varieties of existential risks. To give one example, the deployment of advanced AI-enhanced surveillance capabilities[32] — including automatic hacking, geospatial sensing, advanced data analysis capabilities, and autonomous drone deployment — may greatly strengthen global efforts to protect against "rogue" actors engineering a pandemic ("preventing existential risk"). It may also offer very accurate targeting and repression information to a totalitarian regimes,[33] particularly those with separate access to nanotechnological weapons ("creating a new existential risk"). Finally, the increased strategic transparency of such AI systems might disrupt existing nuclear deterrence stability, by rendering vulnerable previously "secure" strategic assets ("lowering the threshold to existential risk").[34]

Finally, many "non-catastrophic" trends engendered by AI — whether geopolitical disruption, unemployment through automation, widespread automated cyberattacks, or computational propaganda — might resonate to instil a deep technological anxiety or regulatory distrust in global publics. While these trends do not directly lead to catastrophe, they could well be understood as a meta-level existential threat, if they spur rushed and counter-productive regulation at the domestic level, or so degrade conditions for cooperation on the international level that they curtail our collective ability to address not just existential risks deriving from artificial intelligence, but those from other sources (e.g. synthetic biology and climate change), as well.

These brief examples sketch out the broader existential challenges latent within AI research and development at preceding stages or manifesting through different avenues than the signature risk posed by superintelligence. Thus, addressing the existential risk posed by superintelligence is both crucial to avoiding the "adverse outcome", but simultaneously misses the mark in an important sense.

Re-Examining Existential Risks: Hazard, Vulnerability, and Exposure

While Bostrom's leading typology identifies the general area inhabited by existential risks, it provides little guidance for how to differentiate among the diverse risks within that category (the box marked "X"), because these risks are not distinguished according to their source, characteristics, or complexity, but only their impact ("crushing") and scope ("pan-generational").[35]

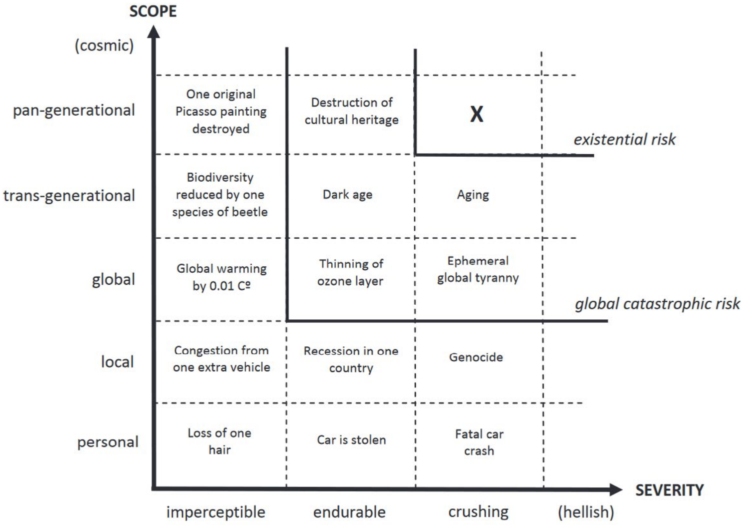


Fig. 1: Qualitative risk categories, indicating the relative position of existential risks. (Reproduced from Bostrom, 2013, p.17).

However, given the range of distinct risks falling within the "X" box — that is, risks that could cause or feed into an eventual terminal and crushing "adverse outcome" for humanity — we suggest it relevant to deconstruct existential risks, and instead consider the broader category of "risks as a function of *hazard, vulnerability* and *exposure*":[36]

Existential Risk = Hazard * Vulnerability * Exposure[37]

Here, *hazard* denotes the external source of peril (which is captured within the prevailing agenda studying existential risks) — the “spark” that threatens the pan-generational/crushing harm.

Vulnerability denotes propensities or weaknesses inherent within human social, political, economic or legal systems, that increase the likelihood of humanity succumbing to pressures or challenges that threaten existential outcomes.

Finally, *exposure* denotes the “reaction surface” — the number, scope, and nature of the interface between the hazard and the vulnerability.

Thus, a hazard is what kills us, and a vulnerability is how we die. Exposure is the interface or medium between what kills us, and how we die. To take an example from disaster studies, a major earthquake only becomes a risk if the built, social or institutional environment can be destabilised during earthquakes of the threatened magnitude (“is vulnerable to”), *and* if such an environment is located in (‘exposed to”) an earthquake zone. Thus, vulnerability and exposure refer to two different aspects of the affected system: how it breaks, and how it intersects with a given hazard’s operating space or pathways of impact.

As a species of Global Catastrophic Risks, the study of existential risks is often conflated with, and perhaps even collapsed into, the identification and mitigation of existential hazards. Where attention is paid to issues of vulnerability and exposure, these are often identified in light of an existential hazard. One of the leading sources and reference points in the field symptomatically organises the field as a collection of existential hazards.[38] A caveat applies for a small subset of hazards of such enormous magnitude that it renders mitigation strategies focusing upon vulnerability and exposure less relevant, or perhaps even irrelevant. The paragon might be the scenarios of “simulation collapse”, or a high-energy physics experiment going awry, altering the astronomical vicinity and rendering life untenable.[39] Such extreme hazards constitute the archetype of existential risks as a subset of Global Catastrophic Risks and can only be addressed by managing the hazard head-on, with vulnerability and exposure components relegated to marginal roles:

$$\text{Existential Risk} = \text{\Large Existential Hazard} * \text{\scriptsize Vulnerability} * \text{\scriptsize Exposure}$$

Thus, our claim is not that the field of existential risk research is looking in the wrong places — the emphasis on existential risks has enabled this field to identify a core group of existential hazards which would on their own suffice to bring about the "existentially adverse outcome".

Nonetheless, there are also many other, slower and more intertwined ways in which the world might collapse, without being hit by spectacular hazards. To complement the study of existential risks we can draw upon lessons learnt through historical and anthropological studies of civilisational collapse. Thus, while existential risks concentrate upon clear-cut existential hazards, civilisational collapse research infers influential factors that were involved in trajectories of decline. These studies are beginning to challenge the traditional conceptual framework which set out a cyclical history, wherein civilisations rise and fall, progressing through a predictable pattern of growth, zenith and decline in a gradual manner.[40] In other words, historically civilisational collapses are boring. Diamond refined this model by recognising that civilisational collapse could be a slow and protracted process emerging from complex interactions.[41]

Beyond Hazards: Vulnerability and Exposure

In this chapter, we set out to foreground the other two variables involved in the existential risk equation.

Thus, as noted, "vulnerability" denotes propensities or weaknesses inherent within human social, political, economic or legal systems that increase the likelihood of humanity succumbing to pressures or challenges that threaten existential outcomes.

"Exposure" indicates the nexus between external hazards and internal vulnerabilities: the interface at which the "adverse outcome" precipitates from their interaction. Historical studies of civilisational collapses indicate that even small exogenous shocks can destabilise a vulnerable system.[42] Given this, studying "exposure" is relevant in systematically analysing interaction effects: a cataclysmic hazard interacting with robust and resilient human systems may be survivable, but conversely, at the interstices at which our human technology, institutions or culture are most vulnerable, even minor (initially "non-catastrophic") hazards can be the inflection point that tips these susceptible systems towards trajectories of collapse.[43]

In order to offset the tight coupling between existential risks and existential hazards, we will further dissect the vulnerability and exposure factors introduced in the existential risk calculus. Our proposed taxonomy distinguishes four general categories of vulnerability and exposure (see Table 1).

- Ontological: vulnerability through existing in a given location and time in our universe;[44]
- Passive: vulnerability through lack of action; "indirect" exposure;
- Active: vulnerability because of insufficient/mis-specified action.
- Intentional: vulnerability or exposure knowingly maintained, for that purpose.

Note that for vulnerability, the Passive, Active and Intentional categories correspond to the jurisprudential concepts of "omission" ("failure to act"), "negligence" (action, but with failure to exercise the appropriate care to prevent foreseeable future harm) and "intention" (action with the known purpose to bring about a consequence).

Drawing such distinctions offers the opportunity to be more precise about the features or characteristics which give rise to the existential dimension of the challenge, and thus suggest specific points for targeted intervention, as well as potential failure modes to caution against.

Table 1: The general categories of vulnerability and exposure, used to structure our taxonomies of existential vulnerability and existential exposure.

	Type of Vulnerability (V) Vulnerability by...	Type of Exposure (E) Exposure by...
Ontological (O)	Existence (V-O)	Existence (E-O)

Passive (P)	Omission (V-P)	Indirect link (E-P)
Active (A)	Negligence (V-A)	Direct link (E-A)
Intentional (I)	Intention (V-I)	Intention (E-I)

Below, we combine these categories and their sub-divisions, in twin taxonomies of existential “vulnerabilities” and “exposure”. We also seek to give concrete examples. Obviously, not all of these examples are currently unstudied — indeed, many feature prominently in the existing literature — though in other cases they remain understudied. While this list is, naturally, not comprehensive, we hope that such examples enable researchers in the field of existential risks to locate their research in an overarching framework, as well as facilitating links to established scholarly fields which have studied given issues, without considering their bearing on larger existential risks.

A Taxonomy of Existential “Vulnerability”

Our proposed taxonomy for distinguishing between different manifestations of existential vulnerabilities is summarised in Table 2: note that the salience or tractability of these existential vulnerabilities to law and policy approaches increases as one goes down: ontological vulnerabilities appear (at present) highly intractable to mere law and policy — it would be a vain regulator indeed who would try to legislate against physical laws. However, as one proceeds to passive, active, or intended vulnerabilities, the salience of governance approaches increases.

Table 2: A taxonomy of vulnerabilities which contribute to existential risks.

Category	Description	Sub-Distinction	Examples of Existential Vulnerabilities
V-O. Ontological vulnerability	Vulnerability that is inherent in being, at present		Simulation shutdown; Biological dependence on continuous/frequent energy & resource inputs (including food, water, air, light, ...); Physical dependence on physics integrity; our biochemistry "works" only within a narrow subset of all possible physical laws (rendering us vulnerable to vacuum decay); Biological ageing.
V-P. Passive vulnerability	Vulnerability existing due to the lack of structures in place. [OMISSION]	*Built* (vulnerability because of the lack of availability of a defence)	Lack a of super-volcano warning system (technology does not yet exist — lack of global capacity). Lack of asteroid defence program (existing technology, but not deployed — lack of local capacity at key point);
		Institutional (top-down social vulnerability)	Lack of effective global institutions, as well as crisis management organisation; Lack of global coordination on identifying and addressing existential risks. Lack of public investment in developing critical technologies, e.g. alternate food sources for surviving volcanic winter[45] or refuges for Global Catastrophic Risks.[46]
		Cultural (bottom-up social vulnerability)	Lack of public engagement in confronting existential risks: propensity of public to stereotype/dismiss disaster scenarios ("Terminator headlines"); Lack of (widely shared) concepts and language to express existential vulnerabilities.

Category	Description	Sub-Distinction	Examples of Existential Vulnerabilities
V-A. Active vulnerability	Vulnerabilities existing in spite of / because of the social structures in place [NEGLIGENCE]	*Built* vulnerability	Intrinsic path-dependent vulnerabilities in infrastructure *components*: architectural security deficits in universally used components of global (digital) infrastructures (e.g. Spectre and Meltdown exploits in Intel chips); future geo-engineering projects, such as stratospheric aerosol injection, which could backfire heavily if interrupted temporarily, and which might be disrupted.[47] Intrinsic path-dependent vulnerabilities in infrastructure *configuration*: critical infrastructures (e.g. national electricity grids) are centralised and homogeneous (e.g. rendering society vulnerability to solar flares). More generally: driven by organisational and competitive optimisation. ("Moloch" traps[48]), globalisation homogenises all solutions across the globe, eroding resilience (e.g. proliferation of homogenised monocultures of staple crops creates vulnerabilities to engineered crop diseases).
		Institutional vulnerability	Narrow bureaucratic interest and perverse incentives which lock civilisation into "inadequate equilibria",[49] potentially blocking coordination for known existential risks. Globalised economic and institutional frameworks. Market dependency[50] Overconfident belief in own ability to foresee risks[51] — risk-based governance and incorrect probabilistic approaches which underestimate fat-tail events.

Category	Description	Sub-Distinction	Examples of Existential Vulnerabilities
		Cultural vulnerability	Spread of pandemics caused by culturally determined interactions (e.g. Ebola); Ingrained distrust of governmental authorities / public media undercutting disaster response efforts; Social norms promoting high fertility and unsustainable population growth.[52] Globalised diets and food demand that can only be met by (unsustainable; vulnerable) monocultures. Increasingly homogenous global "monoculture" in practices and ideology creates vulnerabilities, by limiting redundancies and diversity.
V-I. Intended vulnerability	Vulnerability maintained for a direct purpose [INTENTION]		Misaligned, "apocalyptic" AI;[53] Nuclear force posture combining centralisation of launch command authority, with fallible nuclear early warning systems and "launch-on-warning" missile force postures.[54] "Back-doors" or "zero-day-vulnerabilities" in critical infrastructure software, knowingly maintained by intelligence services. Existence of "omnicidal" agents [55] — including religious groups' faith in end-times, e.g. the Rapture or Yawm ad-Dīn.

7.1 Ontological vulnerability

The category of ontological vulnerability denotes intrinsic vulnerabilities associated with human existence. These include the possibility that we inhabit a computer simulation,[56] which might be terminated or altered at any time. More conceptual and basic vulnerabilities — so fundamental that we often would not even consider them as such — include our existence as biological beings that are dependent (potentially more so than other species such as tardigrades) on continuous or relatively uninterrupted inputs of energy and resources (such as food, water, air, light, ...), which renders the human species one comparatively vulnerable to "extinction" events such as a supervolcano — or meteor-induced global winter. On a deeper level yet, all biochemistry is dependent on the existing laws of physics within which it evolved, rendering us acutely and terminally vulnerable to any processes (e.g. vacuum decay) which would profoundly alter these processes. Biological deterioration due to ageing processes or exterior damages might also rank amongst these, although that is conditional on whether or not there exists a physical "hard ceiling" to how far medical senescence research might extend human lifespans and reduce other vulnerabilities.

As these are background conditions at the frontiers of epistemology, we are unlikely to be able to unveil more than a fraction of these vulnerabilities. Also, as inherent features of human existence we have limited abilities to act effectively in this category. Perhaps the most utility we can extract from delimiting ontological vulnerability is to restrict its reach: in other words, to leave this as a residual class of vulnerabilities inherent in existence.

7.2 Vulnerabilities, passive and active; Built, institutional and cultural

Passive vulnerabilities are characterised by inaction: the susceptibility to existential outcomes by virtue of failure to take appropriate measures. Conversely, *active* vulnerabilities arise in association with human activities, as by-products or unintended consequences.

Three cross-cutting sub-distinctions can also be made for both passive and active vulnerabilities: built, cultural, and institutional.

Built vulnerabilities are characterised by our (passive) failure to put into place relevant solutions or defences to existential challenges, or by our (active) failure to repair or correct the extant vulnerabilities in the legacy infrastructures we deploy, or the path-dependent ways we deploy them — even if we have such solutions or repairs at our disposal. Such solutions can in fact include some interventions proposed by the existential risk research agenda, such as an asteroid defence programme or the ability to systematically monitor for supervolcano eruptions;[57] they also cover the active existential risks posed by the technologies which humanity has introduced, but which go unfixed — such as architectural deficiencies creating intractable cybersecurity vulnerabilities in universally used computing chips. Because of the technical nature of engineered vulnerabilities, some of these are perhaps closest to the existing (policy) research agenda of the existential risk community — and at present some may consider that law and policy tools have less of a role to play, other than to coordinate efforts aimed at addressing them.

In contrast, top-down vulnerabilities resulting from suboptimal direction and coordination are captured by our sub-category of *institutional vulnerability*. Here, the line between active and passive is admittedly thin, where recklessness can be the distinguishing feature. Active institutional vulnerability may be characterised by failure to coordinate to address a known risk, such as climate change, or cyclical global economic meltdown. Passive institutional vulnerability may then be understood as directional and coordination failures that limit the scope of knowledge related to existential risks — perhaps an implicit "unwillingness to know", which translates in an unwillingness to fund blue-sky research into charting "unknown unknowns".[58]

Cultural vulnerability encompass the bottom-up societal dimensions, reflecting how certain social practices may affect susceptibility to existential challenges. Active cultural vulnerabilities include customary practices that facilitate the spread of pathogens, increasing susceptibility to pandemics — for example, integrated commercial travel networks and interpersonal greeting rituals which encourage physical proximity or contact. Passive cultural vulnerabilities include the exclusion or ridicule of existential risks from serious discussion in public forums (let alone the halls of power). This increases collective vulnerabilities insofar as the public and policymakers underrate the prospects for existential risks,[59] resulting in further marginalisation.

7.3 Intended vulnerabilities

Intended vulnerabilities are those which are created or retained specifically for that purpose, and within the existing research agenda are reflected in the premises of the "AI risk" or "Apocalyptic AI" movement.[60] Another salient example can, however, be found in nuclear force postures which (in the US context) features centralisation of launch command authority along with a "launch-on-warning" doctrine that relies on input from fallible early launch warning systems.[61] Together, this gives rise to the catastrophic risk of an accidental nuclear war.[62] Yet, far from incidental, this is arguably by design. As the theorist Kenneth Boulding once observed: "if [deterrence] were really stable ... it would cease to deter. If the probability of nuclear weapons going off were zero, they would not deter anybody".[63, 64] The nuclear force knowingly renders itself more vulnerable to catastrophic accidents — sacrificing a degree of safety for the sake of strengthening operational readiness and deterrence. Less dramatic, similar intentional vulnerabilities could emerge from a state intelligence service knowingly holding back back-doors or "zero-day-exploits" which it identifies in critical infrastructure software, in the hope that this may enable more effective cyberattacks against rival states at a later state.

7.4 Existential vulnerability: Mitigation and adaptation strategies

This taxonomy of vulnerabilities can provide concrete suggestions for addressing existential risks. While the categories of ontological and intended vulnerabilities may seem superfluous, their treatment as additional classes allow limited resources to be concentrated into the most tractable areas. Perhaps the main contribution of this taxonomy is to highlight how existential risks need not be active and discernible, in the manner of the "hazards" identified in the field. Instead, many of these risks can be latent, and slow-moving. Moreover, this taxonomy aids in understanding how human activities can impact paths towards "existential outcomes" in several ways: (1) intent: by directly creating technologies which pose existential hazards (i.e. emerging technologies such as AI, nanotechnology and synthetic biology); (2) (negligence) by establishing complex systems for which failure is unavoidable;[65] (3) and by omission, the failure to take steps to confront existential risks.

Beyond merely refining the sources of existential risks, the contribution of this taxonomy lies in creating a roadmap for the study and integration of risks that have not yet received much or consistent attention in the field of existential risks. In doing so, we emphasise a number of existential vulnerabilities, such as global dependency upon a few species of staple crops, or certain types of globalised technologies (e.g. SCADA-based systems in critical infrastructure) that are not commonly recognised as sources or failure points of existential risks.

The study of existential "vulnerability" may suggest that adaptation strategies are preferable to those of mitigation, both because of the inherent complexity underlying both forms of structural vulnerability and because adaptation can now occur simultaneously with mitigation. This is because the vulnerability analysis in effect opens up a parallel system where other trajectories of existential risks are at play. The rough equivalence drawn between traditional existential risks with existential hazards might have the effect of underselling adaptation strategies: it is illogical to conceive of robustness as a defence against the apocalypse, after all. Along with efforts to mitigate or avert existential hazards, however, we can now also plan for adaptation against vulnerabilities. Thus, adaptation strategies are not limited to actions undertaken after "the Fall": instead, they may become rational reactions towards limiting susceptibility to existential risks. In order to explore this potential further, we proceed to examine a taxonomy of exposure.

A Taxonomy of Existential 'Exposure'

As a parallel effort to our taxonomy on existential vulnerabilities, we set out a classification system to differentiate between different forms of exposure. It is worth recalling at this point that we use exposure to express the interface between hazards and vulnerabilities — between what kills us, and how we die. Both hazards and vulnerabilities in isolation remain as potentials: exposure is thus a means of actualising such potential into existential risks.

Such exposure can further be directed towards either the societal or the natural environment. This is about what is directly at risk: our (human) society and the common capabilities and support structures preventing existential risks, or nature and its carrying capacity and resilience to future shocks. Thus, we assert that devastating results for humankind can follow from the collapse of both the societal structures we have built, as well as the natural environments within which these constructed systems are embedded. Again, the distinction allows us to single out different examples and trajectories to build alternative strategies for human survival. As is clear from the examples above, it also draws out lessons for existential outcomes which might not be immediately evident from an analysis of existential hazards alone. For example, when "exposure" is seen from the perspective of the natural environment on which mankind depends, pervasive over-fishing and deforestation, combined with trends in resource demands tracking population growth, may become potentially hazardous activities with the potential to curtail human development in the long run,[66] even if they do not affect most humans directly in the short run.

Table 3: A taxonomy of modes of exposure which contribute to existential risks.

Category	Description	Sub-distinction	Examples of Existential Exposures
E-O. Ontological exposure	Exposure imposed exclusively by existing (as a human on Earth).	–	Outer space events; Super volcanos Potential (hostile) alien lifeforms
E-P. Indirect exposure	Exposure indirectly caused by societal arrangements intended for something else.	Exposure of Society	AI, nuclear power, nanotechnology and synthetic biology. Experimental scientific curiosity
		Exposure of Nature	Global extreme climate change. Over-utilisation of nature: unsustainable fishing or hunting
E-A. Direct exposure	Exposure directly caused by societal structures intended for something else.	Exposure of Society	Lack of political will and institutional inertia leading to "progress traps".[67] War, METI, or cultural sentiment. Unconstrained optimisation processes in society, economics, politics (politicians), which pursue originally legitimate goals but become misaligned as they find ways to achieve these in increasingly perverse ways, or with increasing amounts of externalities (cf. "Moloch"[68]).

Category	Description	Sub-distinction	Examples of Existential Exposures
		Exposure of Nature	Local ecosystem collapse.[69] Urbanisation, agriculture and deforestation
E-I. Intentional	Exposure directly imposed by societal arrangements intended precisely for that purpose.	-	The existence of nuclear and (infectious) biological weapons for strategic purposes such as deterrence. On a more granular level: the retention of deterrent weapons which risk nuclear winter, over "winter-safe" deterrent.[70]

8.1 Ontological exposure

Some exposures are inherent in residing on Earth. Those falling in the category of natural exposure denote existence on Earth itself as the exposure, and include our exposure to Near-Earth Objects (NEO) hitting Earth or supervolcanoes, triggering a protracted volcanic winter. The common denominator underlying this form of exposure is their requirement for measures beyond our present technological capacity to overcome (which, admittedly, can be a moving threshold).

8.2 Indirect and direct exposure

As with the discussion of existential vulnerabilities set out above, the potential of our proposed taxonomy lies in the analysis of indirect and direct exposures. This distinction identifies the exposures that are a direct consequence of human activity, from those that are caused by more complex interactions with other systems.

The theoretical example of high-energy physics research going awry[71] provides an example of societal exposure.[72] A final example of direct exposures are private or unilateral attempts to undertake "Active SETI" — alternately called METI ("Messaging to Extra-Terrestrial Intelligence")[73]—which might expose the rest of mankind to catastrophic risk, should any future contacted alien species prove hostile and capable of interstellar-scale interdiction. These examples illustrate how surfaces of direct exposure (and ways to reduce it) might be overlooked when concentrating upon the hazard alone.

Beyond direct exposures, there is an array of arrangements which jeopardise the human societies that have become dependent on them. This category includes any activity or arrangement which might expose the world to extinction through cascading effects. The development of critical common global infrastructures such as the internet, energy markets, and cultural and scientific harmonisation might be classified as exposures, rather than vulnerabilities, because these reveal new interfaces between hazards and vulnerabilities. Thus, collapse of common infrastructures would trigger cascades which jeopardise civilisational sophistication at the global level,[74] the edifice upon which humanity's long-term potential has been built. Similarly, developments like urbanisation, intensification of agriculture, and even increasing global inequality[75] appear to be factors that create fault lines and further drive exposures to existential vulnerabilities. Here the exposure perspective shows us that only by certain actions or inactions do risks actually materialise fully against civilisation.

8.3 Intentional exposure

Finally, some of these exposures appear to exist intentionally, or at least knowingly or recklessly. The city of New Orleans, Louisiana, provides a microcosm of how dysfunctional behaviour, seen from an existential risk perspective, might be driven by human incentives or rationales operating at different orders. The city is, in design and position, incredibly vulnerable to its natural environment — pinched in between the Mexican Gulf and Lake Pontchartrain and built on the banks of the Mississippi River. Accordingly, some have argued that the most reasonable strategy following Hurricane Katrina would have been to abandon the city permanently.[76] Instead, the affected populations were given incentives to return, with the US government investing billions in the reconstruction of the city, aware that even with improved defences, the city remains unsafe.[77]

Similarly, many populations worldwide, from Tehran and Kathmandu, to San Francisco and Port-au-Prince, persist in known disaster-prone zones, for (legitimate) reasons of culture, history, identity or economy. The purpose of these examples is not to warn the populations of these cities, nor to judge their decision to remain: rather the point is that individuals and societies often make decisions based upon entirely different rationales than a concern for survival. This is an insight that seems to scale to any level of government. In simpler terms, sometimes we choose exposure over safety because of competing considerations, and while this might be productive from a cultural heritage perspective, it remains problematic when seen through the lens of existential risks.

Are Existential Hazards Necessary for Existential Risks?

Having set out taxonomies for differentiating between factors which influence existential risks, the question remains whether all components are necessary to bring about an "adverse outcome". Our initial claim was that existential hazards could be sufficient existential risks, but that they were not necessary to pose such risks.

Returning to the civilisation collapse literature cited above, Ferguson provides a critical insight in contesting the traditional view of cyclical history itself. He posits an alternative conceptual framework by asking the question: "What if history is not cyclical and slow-moving, but arrhythmic?"[78] Continuing, he summarises the perspective we adopt succinctly:

> Civilisations... are highly complex systems, made up of a very large number of interacting components that are asymmetrically organised, so that their construction more closely resembles a Namibian termite mound than an Egyptian pyramid. They operate somewhere between order and disorder — on 'the edge of chaos', in the phrase of computer scientist Christopher Langton. Such systems can appear to operate quite stably for some time, apparently in equilibrium, in reality constantly adapting. But there comes a moment when they "go critical". A slight perturbation can set off a "phase transition" from a benign equilibrium to a crisis — a single grain of sand causes an apparently stable sandcastle to fall in on itself.[79]

Wright echoes this sentiment: "Civilisations often fall quite suddenly — the House of Cards effect — because as they reach full demand on their ecologies, they become highly vulnerable to natural fluctuations".[80] When combined with the observation that hitherto isolated civilisational experiments have now been merged,[81] this raises the spectre that existential risks can coalesce from factors that historically brought about only limited civilisational collapses. Thus, the question we need to pose in this regard is whether vulnerabilities themselves contain the seeds of existential risks.

In this context, we should note that vulnerabilities have often been considered mostly as aggravating factors. As aggravators, then, vulnerabilities are subsidiary considerations restricted to influencing borderline events: where a potential existential hazard impacts humanity, its susceptibility or resilience could determine whether or not that hazard was transmuted into an existential outcome.

In line with vulnerabilities being developed as a separate sphere where existential risks are at play, this section explores the possibility of removing the existential character of the hazard and thus plausibly reducing the calculus to:

$$\text{Existential Risk} = \text{Hazard} * \text{Existential Vulnerability} * \text{Exposure}$$

[and/or]

$$\text{Existential Risk} = \text{Hazard} * \text{Vulnerability} * \text{Existential Exposure}$$

An initial issue is that a catalyst of some sort is required to precipitate the existential risk, because even a system with well-exposed inherent susceptibilities will need something to set it motion. Removing the existential hazard component allows us to explore the possibility that relatively minor occurrences can trigger cascades that emerge as existential risks. But a vulnerability cannot by definition transmute into the existential risk itself absent external input: for this reason, we diminish the stature of "hazard" in the equation to represent our proposition that exogenous shocks need not be the spectacular existential hazards recognised by the study of existential risks. Instead, the external hazards in our revised equation can include insignificant events which go unnoticed (and quite probably involve a large number of minor occurrences).

Contributions and Limitations of Law and Policy Tools for Existential Risks

While our deconstruction of existential risks leads to fairly broad claims, it also provides a few concrete questions and insights. First and foremost, if existential risks can indeed be triggered by non-existential hazards, we need to broaden the scope of investigation in order to draw a more accurate roadmap of the existential risks field: one which can deal with questions of vulnerability and exposure explicitly.

Second, the type of perceived challenge channels the range of appropriate responses which can be developed. While existential hazards may appropriately be met by narrower forms of technical solutions and technologically-oriented mitigation strategies, our broader perspective of existential risks open up other toolboxes to confront existential risks. In particular, social vulnerability and human-driven (anthropogenic) exposure require improved governance and coordination for adaptation strategies. Thus, when we reconstruct existential hazards through the optics of the social systems' inability to withstand them they, per definition, become social phenomena. As noted, many existential risk scholars have recently recognised the importance of reaching out to, and incorporating, law and governance approaches, even where the origin of the existential hazard itself is technological. The critical role of such law and governance approaches should be even more self-evident where the problems in question — the origins of existential vulnerability and exposure — are themselves social, not technological.

This opens up a field for law and governance scholars to work more productively and on an equal footing with technical experts and philosophers. Moreover, this allows for a different set of research questions to be posed as to how we might reduce the vulnerabilities underlying the existential risks against humanity, and our collective exposure to hazards leading to existential outcomes. In doing so, our taxonomy has the potential to elevate relevant aspects of otherwise mundane considerations within politics, economics and society to the plane of existential risks. In garnering this attention, we hope that law and policy tools might be more productively incorporated and deployed as a means to building resilience and robustness. Here, central legal institutions as rights, responsibility and societal relations might in fact contribute substantially to reducing both our vulnerability towards, and exposure to, existential risks.

The obvious limitations of this approach reside in the observation that many contemporary existential hazards, vulnerabilities and exposures are anthropogenic. This raises the spectre of either "iatrogenesis" ("[complications] caused by the healer"), where our attempts at treating a problem accidentally give rise to new, potentially worse ailments. Thus, in our attempt to curtail existential vulnerabilities and exposures, we may inadvertently generate new or different existential risks. Yet, the framing remains critical: the vantage points created in our proposed taxonomy encourages alternative ways of thinking about existential risks and provide different accommodation strategies.

Finally, the perspective provided by existential vulnerabilities might also foster solutions that will be of more general benefit to humanity as tangential effects of efforts taken to reducing our collective vulnerability and exposure to existential risks. While this appears to be of a lower order of concern at first flush, our taxonomy appears to bind existential risks together with phenomena occurring at different levels. In this sense, existential vulnerabilities and exposures may possess fractal characteristics,[82] reflecting the complexity of their constitution. Support for this claim might reside in the scalability of hazards and vulnerabilities in particular: if pedestrian threats can cascade into existential outcomes, for example, then mundane measures might feedback to reinforce humanity against existential risks. Pushing this to its limits, it is possible the seemingly oblique effects of improved governance undertaken to shore up existential vulnerabilities actually end up as one of the very sources of humanity's resilience and robustness against existential outcomes.

Concluding Thoughts

The lessons that we can draw from deconstructing existential risks into hazards, vulnerabilities and exposures can be divided into internal and external lessons for the field of existential risk research.

In terms of the lessons for existential risk research, our taxonomy suggests that we may presently reside in a situation of pervasive risk. In identifying the catalogue of existential hazards looming over humanity, and focusing attention to confronting these challenges, the perception is that the outcome of these efforts is a lowering of the overall probability of an actualised existential risk. If our efforts are not actually achieving this, however (because they do not address vulnerabilities or exposures, only direct hazards), we run the risk of achieving safety that is merely "symbolic": we perceive that we are "all clear" — that we have successfully steered humanity past "existential outcomes" — when we are in fact all the more fragile. Defeating a global pandemic, or securing mankind from nuclear war, would be historic achievements; but they would be hollow ones if we were to succumb to social strife or ecosystem collapse decades later. By proposing alternative paths that lead to existential outcomes, our taxonomy can recalibrate the calculus and reduce the prospect of an existential outcome.

Our taxonomy also provides the groundwork for concrete strategies for meeting the existential challenges revealed by our deconstruction of existential risks. In essence, our taxonomy enables more productive cross-disciplinary cooperation amongst researchers from the existential risk community and various other disciplines, in assessing the dynamics that might lead towards catastrophic or "existentially adverse" outcomes.

This step in itself seems to enhance resilience and robustness by fostering greater variety of policy and governance responses — responses which can move beyond mitigation alone, to extend to adaptation, and which can better anticipate the strengths and weaknesses of governance. Two key limitations latent within such approaches need to be acknowledged. First, that these new perspectives to confronting existential risks import ingrained societal and institutional problems manifest in lower orders of problems. Second, that the additional complexity introduced into the field of existential risks necessarily makes attempts at framing responses more difficult. The payoffs of such a trade-off are open for discussion.

Yet, our deconstruction of existential risks, and the taxonomy we develop to do so, may show promise as tools to help consolidate and expand the field of existential risk research and bring aligned disciplines to bear on the effort to reduce the overall probability of an existential outcome for mankind. But these are early tentative steps to building alternative vantage points from which to examine existential risks: our hope is that the alternative perspectives that these provide will allow researchers in broader fields to bring their expertise to identify trajectories that could lead to humanity's demise, and to devise strategies to obstruct those paths to existential outcomes.

Acknowledgements

The authors thank two anonymous reviewers for driving key elaborations and qualifications in the final version. In addition, Kristian Lauta would like to thank the Centre for the Study of Existential Risk (CSER).

Notes and References

[1] Bostrom, Nick. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9*(1) (2002); Bostrom, Nick. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (2013): 15–31; Bostrom, Nick and Milan M. Ćirković. 'Introduction', *Global Catastrophic Risks*. Oxford University Press (2008); Matheny, Jason G. 'Reducing the risk of human extinction', *Risk Analysis, 27*(5) (1 October 2007): 1335–44. https://doi.org/10.1111/j.1539-6924.2007.00960.x; Rees, Martin J. *Our Final Century: Will Civilisation Survive the Twenty-First Century*? Arrow (2004).

[2] Bostrom, Nick and Milan M. Ćirković. *Global Catastrophic Risks* (1st edition). Oxford University Press (2008); Haggstrom, Olle. *Here Be Dragons: Science, Technology and the Future of Humanity* (1st edition). Oxford University Press (2016); Pamlin, Dennis and Stuart Armstrong. 'Global challenges: 12 Risks that threaten human civilization', *Global Challenges*. Global Challenges Foundation (February 2015). https://api.globalchallenges.org/static/wp-content/uploads/12-Risks-with-infinite-impact.pdf

[3] Sagan, Carl. 'Nuclear war and climatic catastrophe: Some policy implications', *Foreign Affairs, 62*(2) (1983): 257–92. https://doi.org/10.2307/20041818; Asimov, Isaac. *A Choice of Catastrophes: The Disasters That Threaten Our World*. Ballantine Books (1981); Smil, Vaclav. 'The next 50 years: Fatal discontinuities', *Population and Development Review, 31*(2) (1 June 2005): 201–36. https://doi.org/10.1111/j.1728-4457.2005.00063.x; Posner, Richard A. *Catastrophe: Risk and Response*. Oxford University Press (2004); Tegmark, Max and Nick Bostrom. 'How unlikely is a doomsday catastrophe?', *ArXiv:Astro-Ph/0512204* (8 December 2005). http://arxiv.org/abs/astro-ph/0512204; Ord, T., R. Hillerbrand and A. Sandberg. 'Probing the improbable: Methodological challenges for risks with low probabilities and high stakes', *Journal of Risk Research, 13*(2) (2010): 191–205. https://doi.org/10.1080/13669870903126267; Baum, Seth D. and Anthony M. Barrett. 'The most extreme risks: Global catastrophes', in *The Gower Handbook of Extreme Risk*, ed. Vicki Bier. Gower (2016). http://sethbaum.com/ac/2018_Extreme.pdf; Yudkowsky, Eliezer. 'Artificial Intelligence as a positive and negative factor in global risk', in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan

M. Ćirković. Oxford University Press (2008a), 308–45; Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press (2014).

[4] Bostrom (2002).

[5] Including future generations. Broadly speaking, many scholars in this space share an emphasis on the ethical value of far-future humans — for example, Beckstead, Nicholas. *On the Overwhelming Importance of Shaping the Far Future*. Rutgers University (2013), https://rucore.libraries.rutgers.edu/rutgers-lib/40469/ — with some arguing for the absolute prioritisation of reducing human extinction risks (rather than risks that destroy civilisation but would leave some humans alive) on the grounds that these risks would destroy all future generations — Parfit, Derek. *Reasons and Persons*. Oxford University Press (1984): 453–54; Sagan (1983); Ng, Y.-K. 'Should we be very cautious or extremely cautious on measures that may involve our destruction? On the finiteness of our expected welfare', *Social Choice and Welfare, 8*(1) (1991): 79–88; Matheny (2007). Bostrom himself appears to favour the "Maxipok" strategy — "Maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe" (Bostrom, 2013, p.19) — though he takes a slightly broader perspective of mitigating not just "hard" extinction risks but also "Global Catastrophic Risks" which could inflict significant, lasting long-term harm to the trajectory of human civilisation, and which could thereby end up inflicting other categories of existential risks (including "permanent stagnation"; "flawed realisation", or "subsequent ruination" (Bostrom, 2013, p.19).

[6] Bostrom (2002).

[7] Bostrom (2013); Bostrom and Ćirković (2008).

[8] Avin, Shahar et al. 'Classifying global catastrophic risks', *Futures* (23 February 2018), p.1. https://doi.org/10.1016/j.futures.2018.02.001

[9] Though for some work that examines "interaction effects" between different global catastrophes, see Baum, Seth and Anthony Barrett. 'Towards an integrated assessment of global catastrophic risk', *SSRN Scholarly Paper*. Social Science Research Network (2 October 2017). https://papers.ssrn.com/abstract=3046816; Baum, S. D., Jr. Maher and J. Haqq-Misra. 'Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse', *Environmentalist, 33*(1) (2013): 168–80. https://doi.org/10.1007/s10669-012-9429-y

[10] Avin et al. (2018).

[11] Bostrom (2013), p.27.

[12] Brundage, Miles. 'Guide to working in Artificial Intelligence policy and strategy', *80,000 Hours* (13 June 2017). https://80000hours.org/articles/ai-policy-guide/; Sotala, Kaj. 'Cognitive science/psychology as a neglected approach to AI safety — Effective Altruism Forum', *Effective Altruism Forum* (5 June 2017). http://effective-altruism.com/ea/1b3/cognitive_sciencepsychology_as_a_neglected/

[13] Armstrong, Stuart, Nick Bostrom and Carl Shulman. *Racing to the Precipice: A Model of Artificial Intelligence Development* (technical report). Future of Humanity Institute (2013); Shulman, Carl. *Arms Control and Intelligence Explosions*. Bellaterra (2009). https://intelligence.org/files/ArmsControl.pdf

[14] Yudkowsky, Eliezer. 'Cognitive biases potentially affecting judgment of global risks', *Global Catastrophic Risks, 1*(86) (2008b): 13.

[15] Taylor, Peter. 'Catastrophes and insurance', in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković. Oxford University Press (2008): 164–83. https://doi.org/10.1093/oso/9780198570509.003.0012

[16] Posner, Richard A. 'Public policy towards catastrophe', in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković. Oxford University Press (2008): 164–83. https://doi.org/10.1093/oso/9780198570509.003.0013

[17] Baum, Seth D. 'On the promotion of safe and socially beneficial Artificial Intelligence', *AI and Society, 28* (September 2016). https://doi.org/10.1007/s00146-016-0677-0

[18] Millett, Piers and Andrew Snyder-Beattie. 'Existential risk and cost-effective biosecurity', *Health Security, 15*(4) (August 2017): 373–83. https://doi.org/10.1089/hs.2017.0028

[19] Farquhar, Sebastian, Owen Cotton-Barratt and Andrew Snyder-Beattie. 'Pricing externalities to balance public risks and benefits of research', *Health Security, 15*(4) (August 2017): 401–8. https://doi.org/10.1089/hs.2016.0118

[20] Wilson, Grant. 'Minimizing global catastrophic and existential risks from emerging technologies through international law', *Va. Envtl. LJ, 31* (2013): 307.

[21] While not exhaustive, these include: the Centre for the Study of Existential Risk (CSER), the Global Catastrophic Risk Institute (GCRI), the Global Priorities Project (GPP), the Gothenburg Centre for Advanced Studies, the Global Challenges Foundation, and the Future of Humanity Institute (which has recently announced its "Governance of AI Program").

[22] Farquhar, Sebastian et al. *Existential Risk: Diplomacy and Governance*. Global Priorities Project (2017). https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf

[23] Bostrom (2013), p.19.

[24] Bermudez, Jose Luis and Michael S. Pardo. 'Risk, uncertainty, and super-risk', *Notre Dame Journal of Law, Ethics and Public Policy, 29* (2015): 471–96. https://ssrn.com/abstract=2623715

[25] Bostrom (2014); Yudkowsky (2008a).

[26] Cf. Farquhar et al. (2017). For an excellent overview of recent work (both on technical safety as well as strategy and policy) on mitigating existential risks deriving from Artificial Intelligence, see Dawson, Nik. '2017 AI safety literature review and charity comparison', *Effective Altruism Forum* (13 December 2016). http://effective-altruism.com/ea/14w/2017_ai_risk_literature_review_and_charity/; Dawson, Nik. '2018 AI safety literature review and charity comparison', *Effective Altruism Forum* (20 December 2017). http://effective-altruism.com/ea/1iu/2018_ai_safety_literature_review_and_charity/

[27] Chalmers, David J. 'The singularity: A philosophical analysis', *Journal of Consciousness Studies, 17* (2010): 7–65; Good, I. J. 'Speculations concerning the first ultraintelligent machine', in *Advances in Computers* (vol. 6), ed. Franz L. Alt and Moris Rubinoff. Academic Press (1964): 31–88.

[28] Bostrom (2014); Sotala, K. 'How feasible is the rapid development of Artificial Superintelligence?', *Physica Scripta, 92*(11) (2017). https://doi.org/10.1088/1402-4896/aa90e8; Yudkowsky (2008a); Yudkowsky, Eliezer. 'Intelligence explosion microeconomics', *Machine Intelligence Research Institute* (2015). For critiques of the "singularity" claim, see Brooks, Rodney. 'Artificial Intelligence is a tool, not a threat', *Rethink Robotics* (10 November 2014). http://www.rethinkrobotics.com/blog/artificial-intelligence-tool-threat/; Dietterich, Thomas G. and Eric J. Horvitz. 'Rise of concerns about AI: Reflections and directions', *Communications of the ACM, 58*(10) (28 September 2015): 38–40. https://doi.org/10.1145/2770869; Goertzel,

Ben. 'Superintelligence: Fears, promises and potentials: Reflections on Bostrom's "Superintelligence", Yudkowsky's "From AI to Zombies, and Weaver" and Veitas's "Open-Ended Intelligence"', *Journal of Evolution & Technology, 24*(2) (November 2015): 55–87; Plebe, Alessio and Pietro Perconti, 'The Slowdown Hypothesis', in *Singularity Hypotheses* (*The Frontiers Collection*), ed. Amnon H. Eden et al. Springer (2012), pp. 349–65. https://doi.org/10.1007/978-3-642-32560-1_17; Jilk, David J. 'Conceptual-linguistic superintelligence', *Informatica, 41*(4) (27 December 2017). http://www.informatica.si/index.php/informatica/article/view/1875

29 Armstrong, Stuart and Kaj Sotala. 'How we're predicting AI — or failing to', in *Beyond AI: Artificial Dreams*, ed. Jan Romportl et al. University of West Bohemia (2012), pp.52–75. https://intelligence.org/files/PredictingAI.pdf; Armstrong, Stuart and Kaj Sotala. 'How we're predicting AI–or failing to', in *Beyond Artificial Intelligence*. Springer (2015), pp.11–29. http://link.springer.com/chapter/10.1007/978-3-319-09668-1_2; Baum, S. D., Ben Goertzel and Ted G. Goertzel. 'How long until human-level AI? Results from an expert assessment', *Technological Forecasting & Social Change, 78* (2011): 185–95; Brundage, Miles. 'Modeling progress in AI', *ArXiv:1512.05849 [Cs]* (17 December 2015). http://arxiv.org/abs/1512.05849; Grace, Katja et al. 'When will AI exceed human performance? Evidence from AI Experts', *ArXiv:1705.08807 [Cs]* (24 May 2017). http://arxiv.org/abs/1705.08807; Müller, Vincent C. and Nick Bostrom. 'Future progress in Artificial Intelligence: A survey of expert opinion', in *Fundamental Issues of Artificial Intelligence*, ed. Vincent C. Müller. Synthese Library (2016). http://www.nickbostrom.com/papers/survey.pdf

30 Armstrong, Stuart, Anders Sandberg and Nick Bostrom. 'Thinking inside the box: Controlling and using an Oracle AI', *Minds and Machines, 22*(4) (1 November 2012): 299–324. https://doi.org/10.1007/s11023-012-9282-2; Bostrom, Nick. 'The superintelligent will: Motivation and instrumental rationality in advanced artificial agents', *Minds and Machines, 22*(2) (2012): 71–85; Bostrom (2014); Goertzel, Ben and Joel Pitt. 'Nine ways to bias open-source Artificial General Intelligence toward friendliness', in *Intelligence Unbound*, ed. Russell Blackford and Damien Broderick. John Wiley & Sons, Inc (2014), pp.61–89. http://onlinelibrary.wiley.com/doi/10.1002/9781118736302.ch4/summary; Yudkowsky (2008). The field of AI safety is particularly active. For a selection of influential papers, see Amodei, Dario et al. 'Concrete problems in AI safety', *ArXiv:1606.06565 [Cs]* (21 June 2016). http://arxiv.org/abs/1606.06565; Amodei, Dario and Jack Clark. 'Faulty reward functions in the wild', *OpenAI* (blog) (2016). https://openai.com/blog/faulty-reward-functions/; Christiano, Paul et al. 'Deep reinforcement learning from human preferences', *ArXiv:1706.03741 [Stat]* (12 June 2017). http://arxiv.org/abs/1706.03741; Orseau, Laurent and Stuart Armstrong. *Safely Interruptible Agents* (2016); Soares, Nate and Benjamin Fallenstein. *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda* (technical report). Machine Intelligence Research Institute (2014). https://intelligence.org/files/TechnicalAgenda.pdf

31 Rees (2004); Martin, James. *The Meaning of the 21st Century*. Eden Project Books (2006).

32 Notwithstanding interesting developments in "privacy-preserving" homomorphic encryption configurations, for an interesting exploration of which, see Trask, Andrew. *Safe Crime Prediction: Homomorphic Encryption and Deep Learning for More Effective, Less Intrusive Digital Surveillance* (5 June 2017). https://iamtrask.github.io/2017/06/05/homomorphic-surveillance/

33 For a treatment of totalitarianism as a "Global Catastrophic Risk", see Caplan, Bryan. 'The totalitarian threat', in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković. Oxford University Press (2008), pp.504–19.

34 Hambling, D. 'The inescapable net: Unmanned systems in anti-submarine warfare', *BASIC Parliamentary Briefings on Trident Renewal*. British-American Security Information Council (2016). http://www.basicint.org/sites/default/files/BASIC_Hambling_ASW_Feb2016_final_0.pdf; Holmes, James R. 'Sea changes: The future of nuclear deterrence', *Bulletin of the Atomic Scientists, 72*(4) (3 July 2016): 228–33. https://doi.org/10.1080/00963402.2016.1194060; Lieber, Keir A. and Daryl G. Press. 'The new era of counterforce: Technological change and the future of nuclear deterrence', *International Security, 41*(4) (1 April 2017): 9–49. https://doi.org/10.1162/ISEC_a_00273

35 Of course, Bostrom's objective in setting out this typology is merely to differentiate existential risks from the much larger space of unfortunate occurrences.

36 This classification schema is distinct from another recently proposed by Avin et al. (2018), which instead breaks down the analysis of Global Catastrophic Risk scenarios along three different components — (1) a critical system whose safety boundaries are breached by a threat; (2) the mechanisms by which this threat might spread globally to affect the majority of the population, and (3) the manner in which we might fail to prevent or mitigate 1 and 2. While elegant, a discussion of the similarities, differences, and potential (in)commensurability between these two classification taxonomies is out of scope for this present chapter.

37 Expressing the interrelationship of several variables, and not a mathematically valid equation. A way of deconstructing risk, common to disaster studies — see, e.g., Wisner, Benjamin et al. *At Risk: Natural Hazards, People's Vulnerability and Disasters*. Routledge (2004); Perry, Ronald W. 'What is a disaster?', in *Handbook of Disaster Research*, ed. Havidan Rodriguez, Enrico Quarantelli and Russell Dynes. Springer (2007), pp.1–15.

38 Bostrom and Ćirković (2008).

39 Ord, Hillerbrand and Sandberg (2010).

40 Ferguson, Niall. *Civilization: The West and the Rest*. Penguin (2011).

41 Diamond, Jared. *Collapse: How Societies Choose to Fail or Survive*. Penguin (2006).

42 Diamond (2006); Ferguson (2011).

43 Gladwell, Malcolm. *The Tipping Point: How Little Things Can Make a Big Difference*. Abacus (2001). Notably, the resilience of civilisation to catastrophes has had some treatment in the field of global systemic risk — e.g. Baum, Seth D. and Itsuki C. Handoh. 'Integrating the planetary boundaries and Global Catastrophic Risk paradigms', *Ecological Economics, 107* (1 November 2014): 13–21. https://doi.org/10.1016/j.ecolecon.2014.07.024; Centeno, Miguel A. et al. 'The emergence of global systemic risk', *Annual Review of Sociology, 41*(1) (2015): 65–85. https://doi.org/10.1146/annurev-soc-073014-112317; Helbing, Dirk. 'Globally networked risks and how to respond', *Nature, 497*(7447) (2 May 2013): 51–59. https://doi.org/10.1038/nature12047

44 Another possible term could be "anthropic vulnerability".

45 Pearce, J. M. and D. C. Denkenberger. 'Cost-effectiveness of interventions for alternate food to address agricultural catastrophes globally', *International Journal of Disaster Risk Science, 7*(3) (2016): 205–15. https://doi.org/10.1007/s13753-016-0097-2

46 Haqq-Misra, J., S. D. Baum and D. C. Denkenberger. 'Isolated refuges for surviving global catastrophes', *Futures, 72* (2015): 45–56. https://doi.org/10.1016/j.futures.2015.03.009

47 Baum, Seth. 'Is stratospheric geoengineering worth the risk?', *Bulletin of the Atomic*

Scientists (5 June 2015). https://thebulletin.org/stratospheric-geoengineering-worth-risk8396; Baum, Maher and Haqq-Misra (2013).

48 Alexander, Scott. 'Meditations on Moloch', *Slate Star Codex* (blog) (2014). http://slatestarcodex.com/2014/07/30/meditations-on-moloch/

49 Yudkowsky, Eliezer. *Inadequate Equilibria: Where and How Civilizations Get Stuck*. Machine Intelligence Research Institute (2017).

50 Harari, Yuval Noah. *Sapiens: A Brief History of Humankind*. Harper Collins (2015).

51 Burton, Robert A. *On Being Certain: Believing You Are Right Even When You're Not*. St Martin's Griffin (2008).

52 Kuhlemann, Karin. '"Any size population will do?": The fallacy of aiming for stabilization of human numbers', *The Ecological Citizen, 1*(2) (2018): 181–89.

53 Geraci, Robert. *Apocalyptic AI: Visions of Heaven in Robotics, Artificial Intelligence, and Virtual Reality*. Oxford University Press (2010).

54 Borrie, John. 'A limit to safety: Risk, "normal accidents", and nuclear weapons', *ILPI-UNIDIR Vienna Conference Series* (2014). http://www.isn.ethz.ch/Digital-Library/Publications/Detail/?ots591=0c54e3b3-1e9c-be1e-2c24-a6a8c7060233&lng=en&id=186094

55 Torres, Phil. 'Agential risks: A comprehensive introduction', *Journal of Evolution & Technology, 26*(2) (2016). https://doi.org/10.55613/jeet.v26i2.58

56 Bostrom (2002).

57 Denkenberger, David C. and Robert W. Blair Jr. 'Interventions that may prevent or mollify supervolcanic eruptions', *Futures* (2018). https://doi.org/10.1016/j.futures.2018.01.002

58 Rumsfeld, Donald. *Press Conference by US Secretary of Defence, Donald Rumsfeld*. NATO (6 June 2002). http://www.nato.int/docu/speech/2002/s020606g.htm

59 Cognitive biases exacerbate these effects; Kahneman, Daniel. *Thinking, Fast and Slow*. Penguin (2012).

60 Geraci (2010).

61 Borrie (2014); Sagan, Scott D. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton University Press (1993).

62 Barrett, A. M., S. D. Baum and K. R. Hostetler. 'Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia', *Science and Global Security, 21*(2) (2013): 106–33.

63 Boulding, Kenneth. 'Confession of roots', *International Studies Notes, 12* (1986): 32.

64 Indeed, in *Essentials of Post-Cold War Deterrence*, the US Strategic Command recommended a species of potentially risky brinkmanship, arguing that "[t]he fact that some elements may appear to be potentially 'out of control' can be beneficial to creating and reinforcing fears and doubts in the minds of an adversary's decision makers. This essential sense of fear is the working force of deterrence. That the U.S. may become irrational and vindictive if its vital interests are attacked should be part of the national persona we project to all adversaries". Policy Subcommittee of the Strategic Advisory Group (SAG). *Essentials of Post-Cold War Deterrence*. United States Strategic Command (1995). http://www.nukestrat.com/us/stratcom/SAGessentials.PDF

65 Perrow, Charles. *Normal Accidents: Living with High Risk Technologies*. Princeton University Press (2011).

66 Diamond (2006).

67 Wright (2006).

68 Alexander (2014).

69 Kolbert, Elizabeth. *The Sixth Extinction: An Unnatural History*. Bloomsbury (2014).

70 Baum, S. D. 'Winter-safe deterrence: The risk of nuclear winter and its challenge to deterrence', *Contemporary Security Policy, 36*(1) (2015): 123–48. https://doi.org/10.1080/13523260.2015.1012346

71 Cf. Posner (2004); for interesting methodological work on estimating the safety of experiments within particle physics, as a particular case of evaluating risks with extremely low probability but very high stakes, see Ord, Hillerbrand and Sandberg (2010).

72 High-energy physics research can be distinguished from the category of hazards as an example of societal exposure because of the active decision to conduct the relevant experiments. One might also counter that "nature" would be as much exposed to a potential physics disaster as our society would be. While that is certainly the case, we treat it as a case of societal exposure insofar as humankind is impacted by such accidents, rather than by the impact of such accidents on the environment's integrity or carrying capacity.

73 Zaitsev, Alexander. 'Messaging to extra-terrestrial intelligence', *ArXiv:Physics/0610031* (5 October 2006). http://arxiv.org/abs/physics/0610031

74 Wright (2006).

75 Motesharrei, Safa, Jorge Rivas and Eugenia Kalnay. 'Human and nature dynamics (HANDY): Modeling inequality and use of resources in the collapse or sustainability of societies', *Ecological Economics, 101* (2014): 90–102.

76 Richards, Edward P. 'The Hurricane Katrina levee breach litigation: Getting the first geoengineering liability case right essay', *University of Pennsylvania Law Review PENNumbra, 160* (2011): 267–88.

77 Cutter, Susan L. et al. *Hurricane Katrina and the Forgotten Coast of Mississippi*. Cambridge University Press (2014).

78 Ferguson (2011), p.299.

79 Ferguson (2011), p.299–300.

80 Wright (2006), p.130.

81 Harari (2015).

82 Gleick, James. *Chaos: Making a New Science*. Vintage (1997); Johnson, Steven. *Emergence: The Connected Lives of Ants, Brains, Cities, and Software*. Simon and Schuster (2002).

5. Existential Risk, Creativity and Well-Adapted Science

Adrian Currie

Highlights:

- A group's creativity is proportional to the likelihood that its agents will explore a wide solution-space, e.g. testing varied hypotheses, rather than exploiting a narrower, preferred, solution-space, e.g. sticking to hypotheses similar to those that have already been tested. This can arise both due to individual creativity and/or a diversity of prior beliefs.
- In general, science is increasingly promoting conservativism more than creativity and focusing on a relatively small part of the possible research space. One factor behind this is an emerging "economic approach", where scientists seek credit for themselves and their research. This is encouraged by the professionalisation of science, the growth of competitive, peer-reviewed funding, and the emergence of big science. Other factors include the institutionalisation of science, deepening disciplinary divisions, and the dynamics of lab formation.
- While conservative science is not inherently problematic it is ill suited to epistemic situations that demand scientific creativity, like the study of extreme global risk. This often involves unique, unprecedented events; interactions between highly complex, interdependent, "wild" systems; second-order uncertainty; and a high degree of public engagement.

https://doi.org/10.11647/OBP.0360.05

- A research program is well adapted when the standards, incentives and expectations governing investigations are geared towards overcoming the challenges of the relevant epistemic situation. A well-adapted science of existential risk needs to be creative, multi-disciplinary, pluralistic, and opportunistic. Achieving this requires identifying the sources of maladaptation and asking which of these we might do something about.

This chapter was first published as a paper in *Studies in History and Philosophy of Science Part A* in 2019 and sets out a research agenda for making the study of existential risk more well adapted to its epistemic situation that has profoundly impacted the research culture and approach of the Centre of the Study of Existential Risk (CSER). The tension between exploration and exploitation is further explored in a number of chapters, with Chapter 11, for example, embracing a highly exploratory approach, and Chapters 8, 9 and 16 each presenting a different type of reflection on these different modes.

---

1. Introduction

I'm worried that contemporary science is insufficiently creative to handle some of the more extreme, if improbable, risks from emerging technology. To capture my worry, we'll need to consider the *social epistemology* of science.[1] Where traditionally philosophers took the locus of knowledge to be the individual — what should *I* believe given my sensory evidence, say — social epistemologists recognise that epistemic agents are fundamentally social agents. This is crucial for understanding science: the capacity for scientific learning, publishing, and dissemination depends on interconnected networks, databases, and institutions. Whether or not we think the fundamental locus of knowledge, where it lives, is at the individual or group level, in attempting to understand or explain the epistemology of science, the isolated scientist peering into a microscope is an impoverished starting point. To understand science, we should look to the group.

The *economic approach* is increasingly popular in the social epistemology of science.[2] Individual scientists are taken to be incentive governed agents — credit-maximisers — and scientific communities are modelled on this basis, using simple equations or more complex simulations. This perspective allows us to test the robustness of common platitudes about good science: that there ought to be a division of labour,[3] or that information should flow freely,[4] for example. My approach differs but is also complementary. As opposed to using formal models, I'll use a thick description[5] of what I call a scientific endeavour's *epistemic situation*: roughly the conditions of knowledge generation and the challenges facing it.[6] In doing so, I'll draw on insights from the economic approach and — I hope — inform it as well. I'll suggest that the capacity for simple models to inform, guide, and understand scientific practice is amplified by contextualisation in the manner I illustrate. Specifically, the lessons of such models ought to be put in contact with the epistemic situation at hand, and in some contexts fine-grained, specific detail might make a difference. For the social epistemology of science, then, local details matter. Further, philosophers are increasingly concerned with the role of non-epistemic values in science: their role in setting evidential standards, and in distributing epistemic resources.[7] Bringing these two thoughts together, I'll introduce a notion of science being *well adapted*. A research program is well adapted when the standards, incentives and expectations governing investigations are geared towards overcoming the challenges of the relevant epistemic situation. As we'll see, the notion of a well-adapted science helps integrate work both on scientific values and the economic approach.

I'll make my argument by analysing a case which is both urgent and, I'll suggest, challenges science's adaptedness: the study of human species-level threats, or *existential risk*. Paradigm existential risks have a similar profile: they are more-or-less unprecedented, large-scale, complex, and improbable. Understanding such risks (let alone knowing how to mitigate them) requires creative multi-disciplinary work. Communicating such risks — given their Hollywood-blockbuster-potential — requires delicacy. This all makes for a tricky epistemic situation, particularly considering that contemporary science is geared towards the conservative rather than the creative.

My intention, then, is to both inform philosophical reflection on the social epistemology of science, and to plead for a better-adapted science of existential risk. I'll begin with an account of scientific creativity which philosophers following the economic approach will find familiar. I'll adapt recent work on creativity in human development to distinguish two modes of problem-solving, and suggest this can be adapted to understand the creativity of scientific communities. In short, this consists of *cold searches*, where close locations within a solution-space are methodically examined, and *hot searches*, involving leaps across solution-space. For my purposes, the former counts as conservative and the latter as creative. With this in place, I'll argue that the social organisation of science encourages cold searches. I'll then turn to existential risk, describing the epistemic situation at hand, and arguing that such a situation demands hot searches. I'll close with a discussion of my two themes. First, science's incentive structures should be well-adapted to local conditions, suggesting that the economic approach is most informative when contextualised to an epistemic situation. Second, I'll consider the challenges facing a science of existential risk: what does a well-adapted science of low-probability, high-impact events look like? This last discussion is more a promissory note then a set of concrete policy suggestions: my aim is to clearly identify the challenges faced, and thus set and motivate a research program into how those challenges might be met or mitigated.

2. Scientific Creativity

'Creativity' is polysemous,[8] and I won't attempt an exhaustive account of its guises. Rather, I'll provide a definition which is (1) grounded in current science, (2) lends itself to social epistemology, and (3) illuminates contemporary philosophy of science.[9] Drawing on recent work in developmental psychology, itself taking cues from Artificial Intelligence, I'll characterise scientific creativity in terms of how a solution space is explored. I'll distinguish between "hot" and "cold" searches. These terms are inspired by thermodynamics, and refer, metaphorically, to the kinetic energy of a molecule (a scientist) or a collection of molecules (a research community). A hot molecule, with plenty of stored energy, will move through a space in erratic bounds. A cold molecule will move more slowly. How the metaphor plays out should become obvious.

Gopnik et al. explore problem-solving across different life history stages using a simple experimental paradigm.[10] The underlying thought is that *H. sapiens*' distinctive long childhood might serve to facilitate an individual's adaptation to the various environments it might be confronted with. Because different cognitive powers are better for learning about an environment, and for successful behaviour within it, an early "exploratory" phase could be followed by a less flexible, but better-adapted "exploitative" phase. As they say:

> ...there may be a developmental trade-off between cognitive abilities that allow organisms to learn the structure of a new physical or social environment, abilities that are characteristic of children, and the more adult abilities that allow skilled action in a familiar environment.[11]

Gopnik et al. draw a connection between this cognitive trade-off and one from Artificial Intelligence — between *exploration* and *exploitation*.

> Reinforcement learning algorithms make an important distinction between periods of exploration, in which the system gathers information about potential actions and outcomes, and exploitation, in which information gathering is replaced by taking the actions most likely to maximize reward.[12]

As an individual (artificial or otherwise) adapts to her environment, a trade-off must be struck between learning the lay of the land — exploring the space — and making use of that knowledge — coming to efficient solutions.[13] Gopnik et al. provide experimental evidence that although younger children are typically outperformed by adults when it comes to efficiency, in circumstances where solutions are based on unusual patterns of reasoning, children outdo adults.

One way of fleshing out hot and cold searches appeals to Bayesian priors. Bayesian agents have priors which determine their credence in propositions. Take the numerical sequence 1, 3, 5, 7... My immediate guess as to the next numeral is "9". This is because I have certain expectations — priors — which lead me to favour certain kinds of patterns over others, even though nothing about the sequence *logically demands* that "9" be the correct answer. [14] My predicting "9" is an example of a *cold* search: given some solution space, I'll probe a relatively small area of solutions. A *hot search* would put less weight on priors — I will be more likely to try something out, even if I haven't tried it before, or even if it has failed in the past. Our priors serve to set expectations across a space of possible solutions to a problem. Cold-searching agents will methodically exhaust their local solution-space; hot-searching agents will "jump" about the landscape. From this notion we can define an agent's creativity as follows:

An agent's creativity is proportional to the probability of that agent attempting a distant solution, where a solution's distance is indexed to the agent's priors.

So, creative agents are more likely to attempt solutions at greater distances, while conservative agents will exhaust their local space: the former will be more adventurous.[15] Understanding science, however, requires a grip on community-level processes. It may be that cold-searching individuals nonetheless amount to a creative population. If, for instance, they cluster in widely spaced groups, then despite the relative conservativeness of each agent, wide areas of solution space might be explored. It may also be possible that hot-searching agents don't lead to creativity at the group level. If, for instance, information about previous searches i lost, large jumps in solution space might not lead to the accumulation of information at the group-level. So, we should distinguish between creativity at the individual level and the group level:

A population of agents' creativity is proportional to the likelihood that those agents will explore a wide solution-space.

As noted, being constituted of creative agents is only one way that a population might be creative. Cold-searching agents with widely dispersed initial priors could also lead to a creative population, as could agents using different search algorithms (I'll illustrate further examples in my discussion of the economic approach below). Further, although I've cashed this discussion out in terms of solution-spaces, we could also think of it in terms of evidential sources:[16] instead of creativity informing solutions to a problem, we can think of it as a way of generating evidence pertaining to some hypothesis.

It is important to contrast being "creative" in my sense from being "exhaustive". An exhaustive search will attempt to cover every (or a large number of) solutions in the space, while a creative search will pick out distant solutions. Both creative and conservative populations may achieve exhaustive searches: the latter by systematically attempting options in a small space before slowly expanding; the former by trying wide-ranging solutions and eventually filling in the gaps. In principle, both may be exhaustive, but the search-patterns by which this is achieved would be different.

Creativity in my sense is often — but not necessarily — connected to risk. Insofar as hot searches are more costly, the strategy can involve more risk to the community or individual adopting that strategy. However, this connection is sensitive to epistemic situation. As we'll see in the next section, if we have very little idea of a landscape's topography — if we've as much chance of being located near good solutions as we are bad solutions — then wide-ranging searches could be as risky (or less risky!) than more conservative ones. Again, the risk involved, and the cost of hot or cold searching depends crucially on the local details. This is why "thick descriptions" of epistemic situations are necessary.

So, I've discussed a notion of creativity, grounded in work in AI and developmental psychology, which lends itself to thinking about communities. Again, the account is limited. It doesn't lend itself to understanding what we might call *ingenuity*: some creative individuals have well-trained priors about what tricks and solutions might work in rather outlandish scenarios. There's a difference between a creative search and a chaotic search, and my account is not obviously sensitive to this difference (although considering how agents update their priors might help us here).[17] It also doesn't make room for the creativity involved in cold searches. However, the account is suitable for our purposes here. As we'll see in the next section, it is creativity in this sense which contemporary science is ill equipped to promote, and it is this kind of creativity which scientific study of existential risk requires. I'll next provide a short illustration before connecting my account to the economic approach.

For over a century, dinosaur systematics was founded on a central division between the *Ornithischia* ("bird hipped") and the *Saurischia* ("lizard hipped") dinosaurs: the two groups were taken to go their own evolutionary ways sometime in the mid-late Triassic. This division didn't simply matter for museum displays and documentaries, but shaped questions about the evolution and radiation of dinosaurs in the late Triassic and early Jurassic, as well as the taxonomic allegiances of fossil taxa from near the base of the dinosaur tree. Dinosaurs from the

Triassic were categorized either into *Ornithischia* or *Saurischia* depending on diagnostic characters related to their hip morphology.

In 2017, Baron et al. published a phylogenetic analysis which challenged this received wisdom. In short, by undertaking a wide character reanalysis they generated a phylogenetic tree that drew the basic phylogenetic division in a very different way.[18] Langer et al.'s critical response failed to re-establish the old order.[19] As Baron et al. say in their discussion of that response:

> [the study] results in recovery of the 'traditional' topology, although with less resolution and very weak support; their result is statistically indistinguishable from the possibility that our topology provides a better explanation of the data.[20]

The upshot of all this is not that we should maintain the old order, nor embrace the new: rather, how the base of the dinosaur tree looks is up for grabs. This is an example of a shift from a colder to a warmer temperature science, from conservatism to creativity. Why? Previously, the space of pursuable hypotheses about dinosaur phylogeny, taxonomic membership, and dispersal, was constrained within the traditional picture. Now, it is not: and this opens the door to a much wider set of analyses, hypotheses and interpretations. Where both individual priors and community norms once constrained searches within the *Ornithischian/Saurischian* phylogeny, these have been relaxed, leading to hotter searches, and thus a more creative science of early dinosaurs.

This is just one way in which the creativity (or potential creativity) of a science might increase. There, the undermining of a hypothesis opened up previously uninhabited solution space. However, that is not the only possible route: as we'll see, the social structures of science are likely a source of conservatism, and thus changing these could have similar effects. Now, to link this discussion to the economic approach.

As mentioned in the introduction, the economic approach coopts some of the tools and assumptions of economics (as well as evolutionary biology) to think about scientific communities. Tools such as analytic models and simulations are used and, often at least, scientists are treated as credit maximisers. I'll quickly sketch two popular approaches and relate them to my account of creativity.

First, some philosophers have adapted "bandit models" to examine scientific diversity and communication. In these models, agents pick between two possible options (each representing an arm of a "bandit" gambling machine), where each option has a fixed probability of a fixed pay-off. Typically, one arm is "correct" insofar as it has a higher payoff, a higher probability of payoff, or both. However, agents do not know which arm is better: rather, they have credences which are based upon previous attempts. The model can be used to represent the trade-off between exploration and exploitation discussed above. Upon getting a good result, at what point should an agent simply focus on the lever she considers "the best"? More "creative" agents will take longer to settle on a lever, while less creative — conservative agents — will stick to the lever which gives them the best rewards more quickly. The price of creativity is potentially lower efficiency, while the price of conservativeness is increased possibility of getting stuck on a crappy lever. Kevin Zollman adds network dynamics to the model, in order to test whether open information is always beneficial in science: is it a good idea for each agent to be aware of each other's previous attempts?[21] This illustrates how a population might be creative even if the agents within it are not. By limiting an agent's "vision" — through which other agents' attempts can be seen — the population as a whole becomes more likely to cover more solution space. Of course, in such models "solution space" is small, involving only two options. Landscape models provide a wider perspective.

Epistemic landscape models consist of a three-dimensional space. The X and Y dimensions form a grid. In Weisberg and Muldoon's original formulation, X-Y coordinates designate a research approach to some topic.[22] Say, Gopnik's lab is interested in problem-solving in child development, and one location on the X-Y grid could represent the experimental paradigm they adopt. The Z axis is a series of values which add a topography to the landscape. This axis represents *significance*, with peaks representing important findings. Agents are randomly placed on the landscape and, in effect, explore it attempting to find the peaks. Here, instead of having priors as in bandit models, agent-behaviour is determined by algorithmic instructions. "Followers" will prefer already-explored paths, while "mavericks" will prefer unexplored paths. These models are most often used to understand the division of labour within science: is a population with some proportion of mavericks and followers better than a homogenous one? "Better" is, as in bandit models, understood roughly in terms of the trade-off between exploration and exploitation. A population with too many followers might find itself trapped on a local optima, while maverick strategies are likely costly.[23] Again, we can generate population-level creativity without creative agents. In most landscape models, agents' movement is restricted to their local neighbourhood, and so mavericks cannot "jump" to new locations. Maverick behaviour, however, means that they — and thus the population overall — are likely to explore more space.

Although most uses of landscape models focus on locating peaks, Weisberg and Muldoon note that we can also be interested in maximising the amount of explored space as well. In some cases (I'll suggest one below) we might be more interested in exploring the space of research than finding the "significance peaks". This is particularly the case in more rugged landscapes (that is, landscapes with many peaks). Recently, Michaels, Strevens and Weisberg have modelled various features which lead to "herding" in epistemic communities — which roughly corresponds to cold searching — and have optimistic things to say about potential interventions which encourage hotter science. I take the forthcoming discussion to be complementary.[24]

3. Why Science Is Conservative

I've defined scientific creativity in terms of hot searches, and shown how this discussion fits with recent work in the formal social epistemology of science. That work — necessarily, and not necessarily problematically — abstracts from the actual conditions of scientific investigations. This limits what such models can achieve on their own: they can perhaps test some claims made about science (that, for instance, information-sharing is always good for scientific productivity) and can perhaps motivate more applied questions (that, for instance, a certain amount of population-level creativity can matter). However, how to bring about outcomes, and whether current scientific structures reflect creativity or not, is likely beyond the scope of such models. Here, I will provide an informal argument to the effect that science generates conservative populations. That is, cold searches are promoted at the expense of hot searches. This will matter crucially in the next section when I argue that in some contexts, the study of existential risk in particular, creative scientific populations are necessary for progress. Note that here I won't be distinguishing between more local and incidental — contingent — features of scientific communities, and those which are more general and are perhaps likely to arise in any knowledge-generating practice. This distinction is often critical for understanding science's historical development, and understanding how we might shape current science, but is unnecessary for my project at this stage. My discussion expands upon Kyle Stanford's argument to a similar effect.[25] Stanford doesn't explicitly use my account of creativity, but his discussion is readily adapted.[26]

Stanford, then, emphasises contemporary science's conservativeness. In essence, he acknowledges that science is productive, but argues that this productivity isn't directed towards creativity. In my parlance, scientific incentives encourage cold rather than hot searches. Stanford's approach is based on a comparison between science prior to the 19th century and how it is now, drawing on work in the history and philosophy of science fields, such as Rudwick (1982) and Shapin:[27]

> ...the professionalization of science in the middle decades of the nineteenth century, the shift to state support of academic science through peer-reviewed proposals for particular research projects following World War II, and the ongoing acceleration and expansion of so-called 'Big Science' have served to reduce not only the incentives but also the freedom scientists have to pursue research that challenges existing theoretical orthodoxy or seeks to develop fundamental theoretical innovation.[28]

These features lead to scientific research focusing on a relatively small part of possible research space. I'll summarise and expand upon Stanford's points, before tying this picture of contemporary scientific communities back to my account of creativity. As encapsulated in the above quote, Stanford focuses on three contrasts: the professionalisation of science, the introduction of peer review, and the emergence of big science.

(1) Professionalisation led to scientists' ongoing work being dependent on the approval of their peers. Peer-approval determines what work is interesting, legitimate or significant — and this tends towards consensus forming regarding those matters. That is, the community becomes relatively

homogenous regarding research programs, approaches and perspectives. Although the professionalisation of science meant that science was open to input from a more diverse range of individuals (those from lower social classes, for instance), in order to play the game, one has to buy into the theoretical underpinnings which the players commit to.

(2) The advent of competitive, peer-reviewed funding after World War II amped up conservative inertia. This gave peers more influence over what scientific work was done, thus further empowering consensus on the legitimacy or otherwise of scientific questions, approaches, and so forth. Additionally, these funding sources were often highly centralised and small in number. This encouraged less diversity in what got funded, and encouraged the funding of large projects. Further, getting funding typically requires explicit, pre-decided goals for research with likely epistemic dividends. This discourages open-ended, exploratory research[29] and makes innovative, risky proposals unlikely to be green-lighted.

(3) The advent of "big science" led to a further centralisation and stratification of scientific communities. Producing data for enormous databases requires standardisation.[30] It also more or less necessitates increasingly hierarchical structures in labs wherein scholars are unable to direct research until they are deep into their careers. This further leads to conservative approaches from principle investigators themselves:

> [big science] motivates further intellectual conservatism on the part of advisors and mentors themselves, as a PI who elects to pursue a genuinely revolutionary, transformative, or theoretically iconoclastic research program is more likely to provoke skepticism from a granting agencies' program managers or review committees must now be willing to risk not only her own scientific fortunes but also those of the small army of less well-situated scientific workers whose careers presently depend upon her own.[31]

A useful way of capturing these forces for inertia, implied by the above quote, returns us to the economic approach. Recall that by this approach we should consider scientists as credit-maximising agents. That is, they act not to maximise their own (or the community's) knowledge gains, but their own credit gains. Insofar as scientific credit encourages conservatism, so scientists are led to conservative research agendas. Foster et al. summarise the overall point:

> Scientists "take a position" by pursuing *particular* research problems selected from the space of all those possible. These concrete actions are guided by the interplay between scientists' positions in the field and their *habitus*: acquired systems of taste, dispositions, and expectations. At stake are recognition by fellow scientists, other currencies for which recognition can be traded, and an improved position in the field.[32]

1–3 above provide features of scientific communities which make hot searches risky: they're unlikely to be funded, likely to be treated with suspicion by peers, and unlikely to be supported by large laboratories. This makes such approaches unattractive. Further, the centralisation and increased hierarchy in science likely amplifies the "Matthew Effect". Roughly, by this effect those scientists already in eminent positions will in virtue of that accumulate still further credit:

> ... eminent scientist get disproportionately greater credit for their contributions to science while relatively unknown scientists tend to get disproportionately little credit for comparable contributions.[33]

This clustering of credit potentially undermines the capacity of scientists or labs which are younger, or lesser known, or who are underprivileged (in the third world, for instance) — all potentially diversity-increasing and thus hot-searching — to get noticed, funded, cited, built upon and so on.[34] It also further puts control of scientific output to a small group. Add to this that science is a crowded marketplace: given the small number of available positions, when scientists decide which labs to join, which jobs to apply for or take, which directions to specialise in, which funding to apply for, and so forth, they are making risky bets about which directions will be successful. In a recent editorial, *Nature* reported on a survey of over 5000 early-career scientists. Of these, three quarters intended to pursue careers in science, even though "... only three or four in every hundred PhD students in the United Kingdom will have a permanent staff position at a university. It's only a little better in the United States". In such an environment, it is unlikely that individuals will make things harder for themselves by attempting revolutionary hot searches.

A further force for inertia harkens again back to Thomas Kuhn's emphasis on institutionalised learning. The process of becoming a scientist — transitioning from undergraduate, to graduate, to post-doc, and so forth — doesn't simply involve learning the skills, techniques and theories relevant to that discipline. It also involves taking on a tacit set of expectations about what research questions, approaches, and answers are legitimate and interesting, what Foster et al. above called "habitus". This process of institutionalisation likely promotes scientific productivity: after all, it allows easier communication, and grants the community a common epistemic purpose. However, it also makes it less likely that scientists will be able to overcome their tacit expectations and think revolutionary thoughts.

In addition to these more-or-less tacit expectations within a discipline, in many cases explicit standards for publication become norms. The most obvious of these is the use of p-values to set lower bars for publication across many statistical sciences, and rules against publishing negative results.[35] At least two upshots are relevant here: first, of the research that is done, only some will see the light of day; second, scientists will direct their research efforts towards questions which are more likely to provide results deemed significant by those standards.

Disciplinary boundaries also promote inertia. Insofar as certain disciplinary techniques and research questions constrain researchers to particular parts of research space, without interdisciplinary work much total space will remain unoccupied. And despite claims to the contrary, interdisciplinarity is often discouraged:[36] in addition to the difficulty involved in integrating work from different backgrounds, various gatekeeping processes make its success unlikely.[37] For instance, typically a discipline's most well-regarded journals are in the business of publishing papers concerning the core business of that discipline. This means that interdisciplinary work is published in less-well regarded journals, thus making them less high-profile, and thus less advantageous for career progress and attracting funding. Generally speaking, as the fundamental institutional unit of universities are by-and-large departments of particular disciplines, work on "core" areas of those disciplines is likely to be encouraged. This all adds up to the extra effort required to integrate with people of other disciplines actually hurting one's career.

These forces for inertia are further reflected in, and reinforced by, the behaviours of scientists themselves. Scientists are often suspicious of non-mainstream investigations (or simply investigations outside of their own specialisation), and often engage in informal gate-keeping behaviours. Being marked out as a maverick, or an odd-ball *vis-à-vis* one's scientific endeavours, serves to further isolate potentially revolutionary scientists from the community, thus decreasing their chances of being invited to conferences, being published, being funded, and so on. Such behaviours create what Huw Price has called *reputation traps*: even casual, open-minded consideration of radical scientific ideas can undermine the good name of a once-respectable scientist.[38] A final potential source of inertia comes from the dynamics of lab formation. Cailin O'Connor presents formal modelling results to demonstrate that, under some conditions, a kind of group selection at the lab level can drive conservativeness.[39] In particular, if successful strategies underwriting cold searches are more "heritable" between labs than those underwriting hot-searches (which I think is *prima-facie* likely), then cold-searching is likely to propagate.

This all adds up to a community which pools or herds: a community specialised in cold-searching. Given the institutional, behavioural and tacit forces for conformity, and given the high-risk bets involved in building scientific careers, we should expect scientists to "play it safe" — that is, to choose research paths which are likely to be respected in the community, more likely to provide epistemic dividends, and so forth. That is to say, we should expect modern science to not be creative — to encourage cold searches.[40] This in itself isn't necessarily problematic. Indeed, there are likely to be circumstances where cold searches are just what we want. However, if there are circumstances where a more creative — revolutionary — science is what we need, then modern science is ill equipped to provide it. In the next section, I'll argue that such circumstances exist.

4. The Epistemic Situation of Existential Risk

I've thus far provided an account of scientific creativity as well as reason to think that, most of the time at least, scientific incentive structures do not encourage creativity. In my view, nothing negative follows directly from this. It is only when an epistemic situation demands a creative approach that conservative incentives are problematic: for all I've said thus far cold-searching might be the best strategy for the majority of cases. In this section, then, I want to provide a case study where, I'll argue, creativity is demanded: the study of existential risk. I won't claim that the study of existential risk is unique or distinctive (far from it!); rather, I take it to be a relatively clear example of the kind of epistemic situation which demands a creative science (I'll note caveats as we go). Moreover, as an emerging discipline, characterising it at this stage could encourage reflection on how we should conceive of that work and how it ought to best be practised and shaped. Keep the main point in mind: I've articulated a notion of creativity linked to hot-searching, argued that science doesn't encourage hot-searching, and will now provide an example of an epistemic situation in which hot-searching is called for.

At base, an *existential risk* (X-risk) is a threat to some thing's existence. I take a personal existential risk when I cross the road, and our species takes one when it amasses nuclear weapons.[41] Where many risks — catastrophic risks for instance — are understood in terms of scale (perhaps measured in terms of lives lost, or financial cost), existential risks are indexed to the set of things under that risk. Typically, the study of existential risk focuses on a narrow band of these risks, at the upper end of the bell curve where we meet either human extinction (a species-level threat) or the loss of crucial aspects of civilisation (a culture-level threat). [42] Although the sources of many existential risks are not anthropogenic: extra-terrestrial impacts, supervolcanic eruptions, etc..., the focus of X-risk studies are typically risks from emerging technologies such as Artificial Intelligence, advanced genetic engineering technologies, and synthetic biology.[43] At base, our technological capacities are outrunning our capacity to understand, control or predict the consequences of employing those capacities, and as we'll see this creates a distinctive and difficult epistemic situation.

In this section then, I aim to sketch the *epistemic situation* faced by those studying X-risk. An epistemic situation consists in (1) the challenges facing knowledge generation and (2) the resources available in generating knowledge. Different disciplines and studies face different epistemic situations. Experimental biologists can conduct repeated, fine-grained experimental studies of, say, the developmental systems of fruit flies; whereas scientists testing the effects of pharmaceutical treatments rely on random controlled trials. Presumably part of the reason for the latter is the ethical unsuitability of invasive lab-studies on human subjects. Our epistemic resources and challenges are set by the nature of the systems we're studying, as well as the social, technological and ethical terrain they bump up against. This provides the kind of thick description which, I think, facilitates the contextualisation of work from the economic approach.

Here, I'll focus on the challenges facing X-risk, before briefly discussing the kinds of investigative strategies which might meet those challenges. I'll conclude that the epistemic situation faced by X-risk demands a creative science — in part characterised by hot searches. Therein lies the conundrum at this chapter's centre: the social organisation of science discourages hot searches, but a science of X-risk demands them. Note that not all X-risks share the features I'll list, but I think there is sufficient overlap to be able to talk sensibly about a typical epistemic situation facing scientists interested in paradigm X-risks.

It is worth noting that there are non-epistemic grounds for scientists interested in X-risk to move outside of the usual thinking within their more specialised sub-disciplines. First, consider the importance of highlighting safety concerns. Raising red flags about the potential dangers of new technology is an extremely tricky business. Given the highly competitive nature of funding, and the risky bets scientists take in selecting research directions, pointing out potential risks, especially existential ones, requires individual scientists to put out their necks. If a new technology does get a whiff of the illegitimate, new researchers and funding can flee quickly.[44] As such, the same forces which drive epistemic conservatism in science can also dampen the capacities of scientists working within those fields to raise and study safety concerns. Second, the global nature of both X-risk and the potential benefits of the emerging technologies which raise them likely demand that scientific and technological progress be geared towards the needs of the many, not the few. After all, as X-risks are risks for everyone, the potential benefits of technology which might raise their probability shouldn't be narrowly distributed (particularly to a privileged elite).

Above I've focused on how science is epistemically conservative, but the features I've mentioned might also contribute to conservativism regarding the groups whose interests that research represents. Restrictions on minority groups and those from the global south likely limit the capacity of such crucial sciences to be just. Moreover, lack of representation from those quarters likely themselves restrict scientific productivity.[45]

4.1 Uniqueness

In order to build a theory or model of some phenomenon, it is *prima-facie* plausible that we require multiple examples of it. A unique, unprecedented target, then, presents an epistemic challenge: there is insufficient data to have an empirically grounded model of the phenomena.[46] A pertinent difference between some natural X-risks and those with anthropogenic sources is the events' uniqueness. Asteroids, volcanic activity, and so forth, leave geological signals: we can detect patterns of their occurrence, reconstruct their climatic and biological effects, and generally use the past as a guide to the present. Moreover, our species having already survived approximately one hundred thousand years without a natural event knocking us out makes it defensibly plausible that we're safe for the next (say) hundred years from the kind of extinction risks we faced in the past. However, man-made risks are a different ballgame:

> ... our species is introducing entirely new kinds of existential risk — threats we have no track record of surviving. Our longevity as a species therefore offers no strong prior grounds for confident optimism.[47]

Unique, unprecedented events (or possible events), then, present an epistemic challenge due to both a lack of evidence and an inability to infer from previous behaviour, the result being that uncertainty about risk is likely to dominate risk assessments (Bostrom, 2013).

A science of existential risk, then, must adopt techniques and strategies which mitigate a lack of evidence available to construct theories and models of the relevant phenomena.

4.2 'Wild' systems

The systems involved in X-risk scenarios are often unfriendly to systematic scientific understanding. They are what Kirsten Walsh and I have called relatively *wild systems.*[48] A "wild" system is characterised as being (compared to competing systems) high in both "interference" and "noise". The former concerns the interdependence of the system's parts and their effects: it is difficult to determine the causal powers of particular components in systems of high interference. The latter concerns our capacity to isolate a system: it is difficult to predict the behaviour of a target system which is open to erratic shocks from without. Wild systems — those high in interference and noise — are difficult to study because we cannot isolate and examine their components separately, and their behaviour is often irregular due to exogenous effects.

Human-extinction level threats often involve interactions between highly complex, interdependent systems. Consider extreme solar flares.[49] The occurrence of such an event, in addition to killing the roughly half a million people airborne at any one time, would knock out all satellites and temporarily remove the ozone layer. The effects on global trade, transport, health, politics and communication would undoubtedly be catastrophic — but how catastrophic, and how would the various knock-on effects operate? Answering such questions involves understanding not simply the inner working of particular, complex systems, but also how those behaviours would change, and themselves be changed, by their interdependencies with other systems. Both noise and interference will be high under such conditions. It's important to note that X-risks are not necessarily the outcomes of single cataclysmic events, but in many scenarios emerge from cascades of tragedy which, in combination, add up to civilisational collapse or even human extinction.[50]

A science of X-risk, then, must adopt strategies to mitigate the noisy, high-inference nature of the systems they investigate.

4.3 Second-order uncertainty

Considering uniqueness, I pointed out that uncertainty will likely dominate risk calculus pertaining to X-risks. But that is only one aspect of our ignorance: another concerns which possible events should be on our radar in the first place, and which research questions will be fruitful: to draw on the metaphor of an epistemic landscape, we are ignorant of the landscape's topography — whether there are few peaks, or a more rugged landscape — and of its dimensions: we don't know what the possible sources of X-risks are. In other words, where in 4.1 we focused on known unknowns — that is, our uncertainty regarding the likelihood of some risk — an additional and crucial aspect of our epistemic situation concerns unknown unknowns.

The space of X-risk concerns is already broad: from worries about astronomical events like asteroids and solar-flares, to politics (regarding nuclear capacities, say) to more abstract theoretical worries such as those arising from Fermi's paradox.[51] We lack systematic ways of tackling the space of X-risks.[52]

A science of X-risk, then, should be exploratory: ideally, systematic means of identifying possible sources of risks should be sought.

4.4 The public eye

In addition to challenges emerging from the nature of existential risk itself, a crucial part of the epistemic situation at hand concerns interactions between X-risk and the public.

X-risk naturally lends itself to the splashy: human extinction, the dangers of emerging technology, and so forth, make excellent fodder for science fiction and journalism alike. This brings challenges. A science of existential risk — particularly early on — will get a lot of things wrong. And, indeed, given the low probability of many of the events concerned, it will sometimes be hard to tell when it gets things right. This brings with it two conflicting issues. On the one hand, the public or policymakers might take the science too seriously, and act rashly in light of that. But on the other hand, repeated potential "failures" could lead to a loss of faith in the science.

Further, features of human psychology potentially make X-risk tricky to study insofar as any science needs at least some proportion of positive public regard. Jacob Weiner (2016) has argued that existential and other catastrophic risks face a *tragedy of the uncommons*. In these circumstances, the rareness of an event makes it likely to be misunderstood, mismanaged or neglected. Wiener suggests that, in contrast to typical situations, when facing tragedies of the uncommons experts are more likely to want regulative steps than are laypeople.[53] This is because, first, the rarity and unfamiliarity of the events make them "unavailable" to our minds and imaginations. Second, the scale of the events likely leads to "mass-numbing": a psychological effect where an individual's concern for some costs actually decreases as the cost increases. The effect is possibly because "... respondents feel overwhelmed and doubt that their contribution can really make a difference" (72), or because we respond more to named and known individuals than to faceless masses. Third, our legal and other regulatory institutions are likely to be ineffective in the face of catastrophes, as they will likely break down in those scenarios, thus undermining their motivational power.

A science of existential risk, then, must involve delicate communication with the public and policy-makers.

4.5 Existential risk as a crisis discipline

I have discussed a notion of creativity suitable for examination via the economic approach. One way of contextualising such discussions is via a description of an epistemic situation. An investigation's epistemic situation is the sum of the challenges facing knowledge-generation, and the resources available for overcoming those challenges. I've thus far discussed the challenges facing paradigm X-risk investigations. Paradigm X-risks are *unique*, involve *wild systems*, and involve second-order ignorance. Further, they are in the *public eye*, having the potential to generate over-reactions, a loss of faith, and mismanagement. These challenges are not insurmountable: many of them are not unique to X-risk, and so we can take our cue from other research areas.

It is useful to consider X-risk as a *crisis discipline*. In 1985, Michael Soulé developed the latter notion by comparing conservation biology and cancer research. Both disciplines are geared towards a particular outcome (curing cancer, preserving biodiversity) so membership in the crisis discipline turns on possession of a set of scientific expertise related to achieving that outcome. In addition to ecologists, then, conservation biology includes veterinary specialists, experts in land management, and so on. We've already had a hint about the wide variety of disciplines involved in X-risk — indeed, given our second-order ignorance, it's actually unclear which disciplines will matter. In addition to being multi-disciplinary, crisis disciplines are *normative*: X-risk is not simply in the business of describing or explaining low-probability, high impact events, but also in ascertaining how to minimise the occurrence and impact of such events. A final similarity concerns the need to be tolerant of uncertainty:

> A conservation biologist may have to make decisions or recommendations about design and management before he or she is completely comfortable with the theoretical and empirical bases of the analysis.[54]

In addition to having the characteristics of a crisis discipline, uniqueness and second-order ignorance mean that X-risk studies will often occur in evidentially impoverished circumstances. I've analysed similar epistemic situations occurring in "historical sciences" such as paleontology, geology, and archaeology[55] and here, the success of the sciences is best explained by appeal to the speculative, creative nature of their approach (Alison Wylie has made similar arguments concerning archaeology; see Wylie (1999), Chapman and Wylie (2016)).[56]

I've sketched a set of investigative strategies which maximise evidential reach in historical science. Story-telling and scenario-building serve to maximise the empirical links between hypotheses. Historical reconstruction doesn't simply rely on the relationship between contemporary remains — traces — and the past, but on the connections between our hypotheses about the past. Further, such speculation often generates testable hypotheses (Currie, 2017; Currie and Sterelny, 2017): historical scientists are highly creative in my sense. I've characterised historical scientists as "methodological omnivores".[57] Methodological omnivores engage in two distinctive behaviours. First, they construct epistemic tools and models calibrated to local context (as opposed to using general-purpose tools), enabling rich data to be generated. Second, a pluralistic, opportunistic attitude to techniques, methods and research approaches allows a wide range of perspectives, and different types of evidence, to be available. Finally, uniqueness can be mitigated by the use of partial analogies.[58]

Paradigm X-risks are not precisely in the same epistemic situation as conservation biology or paleontology — the tragedy of the uncommons is one difference — but nonetheless the similarities can give us an inkling of the epistemic strategies which a science of X-risk should adopt. The science should be multi-disciplinary, pluralistic, and opportunistic. Such a science meets the criteria for creativity in the sense I discussed in Section 2. Each of these factors involve a community that does not pool, but rather explores solution space widely.

A science of X-risk, then, should be creative.

5. Discussion

A successful science of X-risk will be creative. But, as we've seen, contemporary scientific incentives don't often encourage creativity. Rather, they encourage cold searches. Hence, the properties required for studying X-risk are not promoted in scientific communities. With this in place, I want to (1) characterise this problem abstractly: that science is "badly adapted" for studying X-risk and (2) continue my initial sketch of a well-adapted science of X-risk.

5.1 Well-adapted science

I've given reason to think that the incentive structures governing science are in a sense "maladapted" for some epistemic situations, existential risk in particular. Where that situation calls for creativity, conservatism is improperly encouraged. In this section, I'll characterise the problem abstractly and discuss its relationship with the economic approach on the one hand, and with work on the relationship between social values and science on the other.

The notion of "well adapted" I want to develop concerns whether scientific incentives encourage the kind of work that is appropriate given an epistemic situation:

A scientific community is well adapted to the extent that the incentives of that community promote the attainment of desired research outcomes, given the epistemic situation at hand.

Let's contrast being well adapted in my sense — which is a relationship between a set of incentives and an epistemic situation — and the notion of an individual scientist or community being adapted to a set of incentives. Work in the economic approach is often not sensitive to this difference, and often the focus is more on the latter. Weisberg and Muldoon's landscape models and their descendants explore how different proportions of exploration strategies might be differentially optimal in various landscapes; Zollman's bandit models explore how scientists might learn to adapt to an epistemic situation. Here, an adapted scientist (or community) is one which maximises their returns (in terms of credit or knowledge) given some set of incentives. To be well adapted in my sense, by contrast, requires those incentive structures to be themselves set in order to maximise the epistemic (or other) outputs that we desire given an epistemic situation. For instance, given the nature of X-risk, incentives in the community should encourage creativity. So, one sense of "adaptive" concerns how well scientific behaviours maximise payoffs given a set of incentive structures. Another — mine — concerns how incentive structures might be organised to maximise epistemic outputs. Again, such a distinction is likely implicit in much of the economic approach, but it is useful to make it explicit.

Socially-inclined philosophers of science have argued that decisions about the pursuit worthiness of a scientific investigation or enterprise — what makes that research program a good one to do — turns on more than epistemic significance. Rather, a cost-benefit calculus is required to balance preferences for research outcomes, the efficiency of investigative approaches to those outcomes, as well as budgetary and ethical constraints. Philip Kitcher's approach is perhaps the clearest example.[59] For him, a science is "well ordered" to the extent that which research we pursue is decided by deliberation which approaches the cost-benefit calculus mentioned above.[60] So, discussion of the role of values in science often draws our attention to how the organisation of science itself affects the efficiency of a research program: concerns about scientific organisation and prioritisation are taken as questions of resource distribution. Given a range of possible questions scientists might be asking, by what principles should they direct their efforts? In short, a well-ordered science is one which balances (1) some suitably trained ("tutored") preferences, against (2) the efficiency of particular research programs in meeting those preferences, and (3) the costs (considered in terms of finances, resources, and ethics) of those programs.[61]

Science might be well ordered — that is, it might target the right programs in an efficient manner — but still not be well adapted. Again, science is well adapted not when the scientists themselves adapt to the incentive structures in place (they'll do that well enough without our help) but when the incentive structure itself is conducive to the achievement of the goals we are interested in. The crucial contrast with Kitcher comes in the second and third aspects of his account of well-ordered science. The efficiency and cost of a scientific endeavour is not fixed, but determined in part by the social context in which the endeavour is carried out. And, to some extent, we have control of that social context.

The notion of a well-adapted science, then, brings two discussions into contact. First, considerations of how scientific communities react to epistemic situations. Second, considerations of the role non-epistemic values play in determining significance in science. A well-adapted science is one where the incentive structures are geared towards achieving the values discussed in the latter literature, and can do so in part in virtue of lessons from the former. On my view, understanding when a research program is well adapted involves local, detailed work: thick descriptions of epistemic situations. The models favoured by the economic approach can play a critical role in suggesting and exploring potential interventions and effects.

5.2 A science of existential risk

I've argued that, insofar as science doesn't promote creativity — hot-searching — it is not well set up for investigating existential risk. In the parlance of the last section, science is *badly adapted* to existential risk. However, researching existential risk is desirable: on the reasonable assumption that human extinction is a bad thing, just a little bit of knowledge which might lower the chances of extinction is going to be worth having. The question, then, is: how do we better adapt science to this epistemic situation? The crucial first step, I think, is to identify the sources of the maladaptation, and the second is to ask which of these we might do something about. I take myself to have gone some way towards the first part of this task. The second part, that is, identifying which aspects of the epistemic situation might be intervened on to better promote research outcomes pertaining to existential risks is tricky and, in this chapter at least, above my paygrade. It is worth noting, however, that in the last sub-section I provided an explicit story for how such intervention strategies might be generated. Simplified models, the bread and butter of the economics approach, can create and explore hypotheses pertaining to the causes of conservatism and their possible interventions. For instance, O'Connor's model should give us pause in assuming that increasing competitiveness will select for conservativeness.[62] Combined with thick descriptions of epistemic situations, and perhaps integrated with empirical data (Harnagel, 2019) such models can then motivate trials of said interventions.[63]

In Section 4 I listed a set of factors which make investigation of X-risk require creative strategies. And in Section 3, I listed a set of factors which make science non-creative. We should ask which of these features discussed in Section 3 may be manipulated in such a way as to make a better-adapted science.

My account of creativity involved a partial trade-off between exploration and efficiency. A science of existential risk should be exploratory, but science is geared towards efficiency and, as we've seen, at least some features promote efficiency at the expense of creativity (although I doubt this is a necessary trade-off).[64] Although paradigm X-risks face a particular epistemic situation, these are not unique insofar as there is bountiful overlap with other sciences. Above, I pointed out similarities between X-risk and sciences like paleontology and conservation biology. These sciences might provide inspiration for how to promote study of X-risk. And indeed some interventions have, generally-speaking, begun to be discussed and partly implemented. The National Science Foundation's "transformation" grants explicitly attempt to fund exploratory research. Some scientific journals have adopted alternative publishing standards: *PLOS ONE*'s policy of publishing any result which is judged to be methodologically sound is an example. And there are at least a few instances of alternative funding allocation strategies being trialled.[65] I take my job in this chapter to be making explicit the underlying reasons for wanting to explore these alternatives — particularly in light of X-risk — but exploring the space of solutions must be left for further work. Those challenges are summarised in Table 1.

Table 1. Sources of scientific conservatism.

Property	Increases Conservatism by…
Peer review (in funding/ publishing)	Tying success to pleasing peers. Slowing down funding/publishing process.
Centralised funding	Tendency towards large projects. Tendency towards safe projects.
Monistic publishing standards	Only some results are published. Bias towards research likely to produce those kinds of results.
'Public eye'	Possibility of miscommunication (either public overreaction or loss of faith).
Crowded marketplace	Scientists hunt out the safest bets in picking research directions.
Explicit success criteria in funding	Makes exploratory research difficult to sell.
Disciplinary focus	Interdisciplinary work/publishing detrimental to career (particularly early on).
Informal gate-keeping (gossip etc…)	Reputation traps
Institutionalised teaching	Tacit consensus
Differences in heritability of success between hot and cold labs.	Lab-level selection for conservative strategies

I see this as a first pass at a research agenda targeting the mitigation of conservation-causing features of science. This list is undoubtedly speculative, surely incomplete, and some features might be mischaracterised or misunderstood. But determining this will require further study, some of which might involve the economic approach, as well as the examination of case studies, and the kind of thick descriptions I have used here. And from this, it is plausible that further interventions might be trialled.

6. Conclusion

The essential tension in this chapter is between aspects of science which make it *productive* — efficient — and those which make it *creative*. I doubt there is a clean trade-off between these virtues, but often we do need to decide whether scientific communities ought to be organised to favour productivity or creativity — cold or hot searches — and to what extent. And those decisions, I've suggested, should be made depending upon the epistemic situations those communities face. For X-risk, contemporary science is far too skewed towards productivity. A well-adapted science of X-risk, then, would be tailored towards generating creativity. I've provided a speculative, preliminary list of the sources of conservatism, and these deserve further study both via empirical and theoretical routes. Especially for cases such as X-risk, understanding how to create well-adapted science is urgent. However, whatever interventions we consider will likely be themselves speculative and risky: and these are risks being taken with the livelihood of individual scientists. In light of this, making such trials fair — providing safety nets, for instance — should be considered part of this research program as well.

Acknowledgements

I'm grateful to Darrell Rowbottom, Shahar Avin, Marta Halina, Thi Nguyen and two anonymous referees for very helpful comments on earlier drafts. This chapter was presented at the *Risk and the Culture of Science* workshop in Cambridge, the comments there helped formed the chapter, as did discussion with my colleagues at the Centre for the Study of Existential Risk. This publication was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of Templeton World Charity Foundation.

Notes and References

[1] I take foundational work in the social epistemology of science to include: Dasgupta, Patha and Paul A. David. ‘Toward a new economics of science’, *Research Policy, 23*(5) (1994): 487–521. https://doi.org/10.1016/0048-7333(94)01002-1; Longino, H. E. *The Fate of Knowledge*. Princeton University Press (2002); Solomon, M. Social Empiricism. MIT Press (2001); Kitcher, Philip. *The Advancement of Science*. Oxford University Press (1993); Strevens, Michael. ‘The role of the priority rule in science’, *Journal of Philosophy, 100*(2) (2003): 55–79. https://doi.org/10.5840/jphil2003100224

[2] Muldoon, R. ‘Diversity and the division of cognitive labor’, *Philosophy Compass, 8*(2) (2013): 117–25. https://doi.org/10.1111/phc3.12000. Take my use of “economic” with a pinch of salt: I think this is a useful, if imperfect, term for a set of more-or-less formal approaches to understanding science. Although much of it does take explicit cue from economics, some does not: landscape models, for instance, are adapted from evolutionary biology.

[3] For instance, Alexander, J. M., J. Himmelreich and C. Thompson. ‘Epistemic landscapes, optimal search, and the division of cognitive labor’, *Philosophy of Science, 82*(3) (2015): 424–53. https://doi.org/10.1086/681766; Pöyhönen, S. ‘Value of cognitive diversity in science’, *Synthese* (2016): 1–22. https://doi.org/10.1007/s11229-016-1147-4

[4] For instance, Zollman (2010); Zollman, K. J. ‘Social network structure and the achievement of consensus’, *Politics, Philosophy & Economics, 11*(1) (2012): 26–44. https://doi.org/10.1177/1470594X11416766

[5] A “thick concept” is one which combines both descriptive and normative content, while the content of a thin concept is merely descriptive (compare “overweight” to “fat”, say). By a “thick description”, I mean here a description of scientific practice which is explicitly normative.

[6] For discussion, see Leonelli, S. *Data-Centric Biology: A Philosophical Study*. University of Chicago Press (2016); and Currie, Adrian. *Rock, Bone, and Ruin: An Optimist’s Guide to the Historical Sciences*. Print. Life and Mind: Philosophical Issues in Biology and Psychology (2018).

[7] Douglas, H. *Science, Policy, and the Value-Free Ideal*. University of Pittsburgh Press (2009); Kitcher, Philip. *Science, Truth and Democracy*. Oxford University Press (2001); Kitcher, P. *Science in a Democratic Society*. Prometheus Books (2011).

[8] See, for instance, Gaut, B. *Creativity and Imagination. The Creation of Art* (2003), pp.148–73; Paul, E. S. and S. B. Kaufman (eds.). *The Philosophy of Creativity: New Essays*. Oxford University Press (2014).

[9] In philosophical accounts dealing with creativity more generally, mine has most affinity with Margaret Boden’s concept of “exploratory creativity“ (Boden, M. A. *The Creative Mind: Myths and Mechanisms*. Psychology Press (2004)), and perhaps clashes most directly with Berys Gaut’s arguments that creativity ought to be an agential property (Gaut, 2010) — see the introduction to this issue (Currie, 2019) for further discussion of the relationship between creativity in philosophy and the concept developed here.

[10] Gopnik, A., S. O’Grady and C. G. Lucas et al. ‘Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood’, *Proceedings of the National Academy of Sciences, 114*(30) (2017): 7892–99. https://doi.org/10.1073/pnas.1700811114

11 Gopnik et al. (2017), p.7892.

12 Gopnik et al. (2017), p.7893.

13 See Gopnik, A., C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir and D. Danks. 'A theory of causal learning in children: Causal maps and Bayes nets', *Psychological Review, 111*(1) (2004), 3 https://doi.org/10.1037/0033-295x.111.1.3; and Tenenbaum, J. B., C. Kemp, T. L. Griffiths and N. D. Goodman. 'How to grow a mind: Statistics, structure, and abstraction', *Science, 331*(6022) (2011): 1279–85. https://doi.org/10.1126/science.1192788. It's worth pointing out that exploration of necessity involves some exploitation in many contexts: we should think of these strategies not in terms of whether exploitation or exploration occurs, but the size of the problem-space considered for exploitation.

14 Norton, J. D. 'A material theory of induction', *Philosophy of Science*, *70*(4) (2022): 647–70. https://doi.org/10.1086/378858. As Norton points out, even if we restrict the sequence to following a single, simple rule, underdetermination remains: the sequences could be the odd numbers, or the odd primes starting from one, or the decimal expansion of 359/2645.

15 Here, I don't intend any particular notion of "probability".

16 See: Currie, A. and S. Avin. 'Method pluralism, method mismatch, and method bias', *Philosopher's Imprint, 19*(13) (2019). http://hdl.handle.net/2027/spo.3521354.0019.013

17 To put it in Gaut's terms, it doesn't include creative "flair" (Gaut, B. 'The philosophy of creativity', *Philosophy Compass, 5*(12) (2010), 1034–46. https://doi.org/10.1111/j.1747-9991.2010.00351.x).

18 Baron, M. G., D. B. Norman and P. M. Barrett. 'A new hypothesis of dinosaur relationships and early dinosaur evolution', *Nature, 543*(7646) (2017): 501–6. https://doi.org/10.7934/p2651

19 Langer, M. C., M. D. Ezcurra, O. W. Rauhut et al. 'Untangling the dinosaur family tree', *Nature, 551*(7678) (2017). https://doi.org/10.1038/nature24011

20 Baron, M. G., D. B. Norman and P. M. Barrett. 'Baron et al. reply', *Nature, 551*(7678) (2017), E4.

21 Zollman, K. J. 'The epistemic benefit of transient diversity', *Erkenntnis, 72*(1) (2010), p.17. https://doi.org/10.1007/s10670-009-9194-6

22 Weisberg, M. and R. Muldoon. 'Epistemic landscapes and the division of cognitive labor', *Philosophy of Science, 76*(2) (2009): 225–52. https://doi.org/10.1086/644786

23 Muldoon, R. 'Diversity, rationality, and the division of cognitive labor', in T. Boyer-Kassem, C. Mayo-Wilson and M. Weisberg (eds.). *Scientific Collaboration and Collective Knowledge: New Essays*. Oxford University Press (2017). https://doi.org/10.1093/oso/9780190680534.003.0004 Thoma, J. 'The epistemic division of labor revisited', *Philosophy of Science, 82*(3) (2015): 454–72. https://doi.org/10.1086/681768

24 Strevens, Michael. 'Herding and the quest for credit', *Journal of Economic Methodology, 20*(1) (2013): 19–34. https://doi.org/10.1080/1350178x.2013.774849; Weisberg, M. 'Modeling herding behavior and its risks', *Journal of Economic Methodology, 20*(1) (2013): 6–18. https://doi.org/10.1080/1350178x.2013.774843

25 Stanford, P. K. 'Unconceived alternatives and conservatism in science: The impact of professionalization, peer-review, and big science', *Synthese* (2015): 1–18.

26 Stanford is responding to criticisms of his arguments for scientific anti-realism (see Stanford, P. K. *Exceeding Our Grasp: Science, History, and the Problem of Unconceived*

Alternatives. Oxford University Press (2010). https://doi.org/10.1093/0195174089.001.0001) on the basis of unconceived alternatives (Godfrey-Smith, P. 'Recurrent transient underdetermination and the glass half full', *Philosophical Studies, 137*(1) (2008): 141–48. https://doi.org/10.1007/s11098-007-9172-2; Forber (2008). The connection between his discussion and mine is science's conservatism: if science encourages cold searches, then it is unlikely to discover alternative theories or hypotheses.

27 Shapin, S. *The Scientific Life: A Moral History of a Late Modern Vocation*. University of Chicago Press (2008). https://doi.org/10.7208/chicago/9780226750170.001.0001

28 Stanford, P. K. 'Unconceived alternatives and conservatism in science: The impact of professionalization, peer-review, and Big Science', *Synthese, 196*(10) (2015): 3915–32. https://doi.org/10.1007/s11229-015-0856-4

29 Currie, A. *Rock, Bone and Ruin: An Optimist's Guide to the Historical Sciences*. MIT Press (2018b): Chapter 12.

30 Leonelli, S. 'What counts as scientific data? A relational framework', *Philosophy of Science, 82*(5) (2015): 810–21. https://doi.org/10.1086/684083

31 Stanford (2015).

32 Foster, J. G., A. Rzhetsky and J. A. Evans. 'Tradition and innovation in scientists' research strategies', *American Sociological Review, 80*(5) (2015): 876. https://doi.org/10.1177/0003122415601618

33 Merton, Robert K. 'The Matthew Effect in science', *Science, 159*(3810) (1968): 58. https://doi.org/10.1017/9781108610834.005

34 See, for example: De Cruz, H. 'Prestige bias: An obstacle to a just academic philosophy', *Ergo, an Open Access Journal of Philosophy, 5* (2018). https://doi.org/10.3998/ergo.12405314.0005.010

35 See, for instance: Benjamin, D. J., J. O. Berger, M. Johannesson et al. 'Redefine statistical significance', *Nat Hum Behav, 2*(6–10) (2018). https://doi.org/10.1038/s41562-017-0189-z

36 Bromham, L., R. Dinnage and X. Hua. 'Interdisciplinary research has consistently lower funding success', *Nature, 534*(7609) (2016): 684. https://doi.org/10.1038/nature18315

37 See, for instance, papers in: Mäki, U., A. Walsh and M. F. Pinto (eds.). *Scientific Imperialism: Exploring the Boundaries of Interdisciplinarity*. Routledge (2017);. Forber, P. 'Forever beyond our grasp?' *Biology & Philosophy, 23*(1) (2008): 135–41. https://doi.org/10.1007/s10539-007-9074-x

38 Price, H. 'The cold fusion horizon', *Aeon* (2016). https://aeon.co/essays/why-do-scientists-dismiss-the-possibility-of-cold-fusion

39 O'Connor, C. 'The natural selection of conservative science', *Studies in History and Philosophy of Science Part A, 76* (2019): 24–29. https://doi.org/10.1016/j.shpsa.2018.09.007

40 This is a very general argument, and there will be local exceptions — some rather glaring. Further work needs to be done to establish to what extent efforts akin to the Nobel Prizes, say, promote riskier scientific strategies. I'm no expert on any of the sciences that get Nobels, nor on the prizes themselves, but it would be interesting to consider what effects such high-impact "heroic" awards in fact have on scientific communities (see, for instance: Wagner, C. S., E. Horlings, T. A. Whetsell, P. Mattsson and K. Nordqvist. 'Do Nobel Laureates create prize-winning networks? An analysis

of collaborative research in physiology or medicine', *PLOS ONE, 10*(7) (2015): e0134164. https://doi.org/10.1371/journal.pone.0134164

41 This is an idiosyncratic definition: there is not, so far as I can tell, an agreed-upon account of existential risk. I prefer to distinguish existential from catastrophic risks in terms of whether the risk is indexed to the risk's subject (in this case, the existence of that subject), or to a scale (say, minor to catastrophic). Bostrom defines X-risk as one "... that threatens the premature extinction of Earth-originating intelligent life or the drastic destruction of its potential for desirable future development" (Bostrom, N. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (2013): 15. https://doi.org/10.1111/1758-5899.12002; Bostrom, N. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9*(1) (2002); Torres, P. *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks*. Pitchstone Publishing (2017). I consider this a feature rather than a bug: human-level extinction risks are a motley bunch, and in different contexts different definitions may be more or less useful. I don't see any pressing reason to insist on a single orthodox definition of the domain of X-risk, so long as there is clarity within particular discussions.

42 See, for example: Bostrom, Nick and Milan M. Ćirković. *Global Catastrophic Risks*. Oxford University Press (2008), Baum, S. D. and A. M. Barrett. 'The most extreme risks: Global catastrophes', in Vicki Bier (ed.). *Risk in Extreme Environments: Preparing, Avoiding, Mitigating, and Managing*. Routledge (2018): 174–84.

43 Although distinguishing between human-caused risks and "natural" risks is sometimes useful, it is important to see that such distinctions have a shelf-life. The scale of "natural" risks depends in part on what measures we have taken to understand and mitigate them.

44 See, for instance, my discussion of geoengineering: Currie, A. 'Geoengineering tensions', *Futures, 102* (2018a): 78–88. https://doi.org/10.1016/j.futures.2018.02.002/

45 O'Connor, C. and J. Bruner. 'Dynamics and diversity in epistemic communities', *Erkenntnis, 84*(1) (2017a): 101–19. https://doi.org/10.1007/s10670-017-9950-y

46 Tucker, A. 'Unique events: The underdetermination of explanation', *Erkenntnis* (*1975–*), *48*(1) (1998): 59–80.

47 Bostrom, N. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (2013): 15. https://doi.org/10.1111/1758-5899.12002

48 Currie, A. and K. Walsh. 'Newton on Islandworld: Ontic-driven explanations of scientific method', *Perspectives on Science, 26*(1) (2018): 119–56; Currie, A. and K. Sterelny. 'In defence of story-telling', *Studies in History and Philosophy of Science Part A, 62* (2017): 14–21. https://doi.org/10.1162/posc_a_00270; Currie, A. 'Creativity, conservativeness & the social epistemology of science', *Studies in History and Philosophy of Science Part A, 76* (2019): 1–4. https://doi.org/10.1016/j.shpsa.2018.11.001

49 Isobe, H. *Extreme Solar Flares as a Catastrophic Risk* [presentation]. Cambridge Conference on Existential Risk (2016).

50 Kareiva, P. and V. Carranza. 'Existential risk due to ecosystem collapse: Nature strikes back', *Futures*, *102* (2018): 39–50. https://doi.org/10.1016/j.futures.2018.01.001; Liu, H. Y., K. C. Lauta and M. M. Maas. 'Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures* (2018). https://doi.org/10.1016/j.futures.2018.04.009

51 Miller, J. D. and D. Felton. 'The Fermi paradox, Bayes' rule, and existential risk management', *Futures, 86* (2017): 44–57. https://doi.org/10.1016/j.futures.2016.06.008

52 Although, see: Avin, S., B. C. Wintle, J. Weitzdörfer, S. S. Ó hÉigeartaigh, W. J. Sutherland and M. J. Rees. 'Classifying global catastrophic risks', *Futures*, *102* (2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001

53 Wiener, J. B. 'The tragedy of the uncommons: On the politics of apocalypse', *Global Policy*, *7*(S1) (2016): 67–80. https://doi.org/10.1111/1758-5899.12319

54 Soulé, M. E. 'What is conservation biology?', *BioScience, 35*(11) (1985): 727–34. https://doi.org/10.2307/1310054

55 Currie, A. 'Hot-blooded gluttons: Dependency, coherence, and method in the historical sciences', *The British Journal for the Philosophy of Science, 68*(4) (2016): 929–52. https://doi.org/10.1093/bjps/axw005; Currie, A. 'Marsupial lions and methodological omnivory: Function, success and reconstruction in paleobiology', *Biology & Philosophy, 30*(2) (2015): 187–209. https://doi.org/10.1007/s10539-014-9470-y

56 Wylie, A. 'Rethinking Unity as a "Working Hypothesis" for Philosophy of Science: How Archaeologists Exploit the Disunities of Science', *Perspectives on Science, 7*(3) (1999): 293–317. https://doi.org/10.1162/posc.1999.7.3.293; Chapman, R. and A. Wylie. *Evidential Reasoning in Archaeology*. Bloomsbury Publishing (2016). https://doi.org/10.5040/9781474219167

57 Currie (2018b).

58 Currie (2018b).

59 Kitcher, Philip. *Science, Truth and Democracy*. Oxford University Press (2001); Kitcher, Philip. *Science in a Democratic Society*. Prometheus Books (2011).

60 See also: Cartwright, N. 'Well-ordered science: Evidence for use', *Philosophy of Science, 73*(5) (2006): 981–90. https://doi.org/10.1086/518803

61 In recent work, Kitcher further develops the notion of a well-ordered science, expanding the role of values in determining the significance of research questions, the role of the public in certifying scientific claims, and the importance of disagreement both within and without science. These developments don't affect the contrast with well-adapted science.

62 O'Connor, C. 'The natural selection of conservative science', *Studies in History and Philosophy of Science Part A, 76* (2019): 24–29. https://doi.org/10.1016/j.shpsa.2018.09.007

63 Avin, S. 'Centralized funding and epistemic exploration', *The British Journal for the Philosophy of Science* (2019). https://doi.org/10.1093/bjps/axx059

64 Philosophers interested in the composition of scientific communities often argue that efficiency is achieved by some mixture "maverick" and "follower" strategies (Kitcher, Philip. 'The division of cognitive labor', *Journal of Philosophy, 87* (1990): 5–22. https://doi.org/10.2307/2026796; Weisberg and Muldoon (2009); Thoma (2015), suggesting that there is not a trade-off between creativity and efficiency. This may be right, but it is worth noting that my conception of creativity is not quite equivalent (as, properly speaking, it is a population-level phenomenon) and, as we've seen, such work is not geared towards understanding science's being well-adapted.

65 Avin (2019).

II. METHODS, TOOLS, AND APPROACHES

As we saw in Section 1, the study of extreme global risks, including existential and Global Catastrophic Risk, raises a number of empirical and epistemological challenges. These risks involve unprecedented phenomena and complex systems moving outside of their "normal" operating space; they demand the synthesis of knowledge and expertise from different disciplines and domains; they are rapidly evolving and yet require long-term planning; and they involve all of humanity but are mostly shaped and perceived by a few elite individuals.

There are many ways that the field can and has been responding to these challenges; however, one of the most important of these is in the creation, development, application, and evaluation of methodologies, and this form of innovation has been core to the work of the Centre of the Study of Existential Risk. Without wishing to get into a philosophical discussion about the nature of methodology, some key features of methods include their generality, in the sense that they can be applied in more than one case; formality, in the sense that they follow a set of rules or procedures; publicity, in the sense that they are known both to the producer and consumer of research; and objectivity, in the sense that their results should depend upon more than the pre-existing beliefs, values, or biases of individual researchers. There are many forms of methodology, including those of the natural sciences (for instance, testing and refuting empirical hypotheses), the social sciences (for instance, producing an objective statement of observed social phenomena), the humanities and mathematics (for instance, presenting a clear chain of reasoning that others can follow), and engineering (for instance, breaking down a complex problem into components whose solution is already known or easy to determine).[1]

As a trans-disciplinary field, Existential Risk Studies has benefited from the work of researchers who make use of all of these methodologies,

[1] These are only intended to be vastly oversimplified examples of the kinds of methodologies used in different fields of study and we apologize to the many researchers who feel their work does not fit into this catagorisation.

and others not mentioned above. Alongside this, researchers have also developed a suite of more focused tools, that can be used to study particular issues or phenomena but are not as generalisable as methodologies, and also less tangible ideas we refer to as "approaches", that provide a stimulus to creative thinking about extreme global risk but are either less formal or less objective than methodologies. In the following chapters we review a range of these methods, tools, and approaches, focusing on those that have been developed or influenced by the Centre for the Study of Existential Risk, but also surveying the wider landscape. Our aim is both to help readers to understand these methods, tools, and approaches and how they work and to feel able to consider whether they could help them in their own work. In this respect, many of these chapters talk about the origin and history of different methods, tools, and approaches, who currently utilises them and why, what they do and do not tell us about risks, how they can inform better policy- and decision-making, and whether they are suited to exploring a diverse possibility space or exploiting a more tightly defined problem set and evidence base.

The chapters gathered in this section provide a wonderful overview of many of the methodological developments taking place in relation to the study of extreme global risk, and also point towards some of the core debates that these have provoked. The works here show how empirical assessments have been utilized in tandem with speculative or exploratory tools to study hazards and vulnerabilities as both obdurate facts and contingent potentialities. Given the field's core focus on events, or sequences of events, that are generally of low probability, but extremely high impact, this interplay between examinations of *what is* and experiments with *what might be* remains a necessary and exciting challenge for the development of methods and approaches for study. The chapters gathered in this section provide some insight into how different approaches taken to date have engaged with, and negotiated these interplays.

Alongside these (hopefully productive) tensions, between empiricism and positivism on the one hand and explorations of possibility on the other, these chapters also point to debates over the form, function and value of generalisability of research methods for existential and catastrophic risks. While research on one category of risks and our means of mitigating it might yield insights that can be explored in relation to other categories, the chapters here also remind us that aspirations for a "perfect" science for studying catastrophic risk shouldn't supersede critical reflections on the actually existing differences between different hazards, their drivers, and the factors that condition our specific and uneven vulnerabilities to them. Works in this section illustrate the need for ongoing reflection, mutual learning from alternative methodologies, and an attentiveness to the interactions between risks and vulnerabilities in a complex system, whilst also demonstrating the limitations of one-size-fits-all approaches or methods that homogenise specific cases.

Another tension that several of these chapters highlight is that between the credibility of different methods and the researchers who apply them. Existential Risk Studies is a field of research that has crystallised around a number of elite academic institutions, and researchers in this space can easily find themselves with a ready platform for their research in virtue of this. However, thinking about methods, tools, and approaches challenges us to look past the capacities and status of individual researchers to consider what they are actually doing in their research. Sometimes, emphasizing this point can actually detract from the credibility of particular claims, especially where methodologies highlight subjective opinions or model assumptions that are open to question. This is something that should be welcomed, and yet as Chapter 5, *Existential Risk, Creativity and Well Adapted Science*, points out, Existential Risk Studies exists in an economic paradigm that often privileges what is seen as good for researchers ahead of what may be better for their research. Ensuring that the right methodologies are being deployed, and that they are used in ways that are beneficial for both researchers and their work, is thus very important for the longstanding health of Existential Risk Studies as a field.

Chapter 6, *An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards*, provides a wide-ranging overview and assessment of different methodologies and approaches in Existential Risk Studies, with a focus on those that have been used to quantify the likelihood of existential hazards. These include analytical approaches based in philosophy and mathematics, extrapolation from available data, toy models, fault trees, Bayesian networks, complex systems models, agent-based models, individual judgement, simple surveys, weighted aggregation, enhanced solicitation techniques, and prediction markets. The chapter uses an informal evaluative framework to consider and assess each of these methods for rigour, ability to handle uncertainty, accessibility for researchers with limited resources, and utility for communication and policy purposes. It finds that different methods have very different profiles of advantages and disadvantages, with no clear "winner" for quantifying existential risk. Nevertheless, some methods may be more suitable to certain purposes within Existential Risk Studies and some, especially the Delphi technique and related forms of Structured Expert Solicitation, may deserve a wider application by the community. More importantly, when turning from methodologies in general to their specific implementation by researchers in the field, the authors find that in many cases, claims based on poor implementations of these methods are still frequently invoked by the Existential Risk Studies community, despite the existence of more methodologically robust estimates. This includes the subjective judgement of high-profile researchers within the field who seldom present a clear methodology for how they arrived at their claims, and eye-catching claims by researchers from other fields that may not be given much credence by disciplinary colleagues. The chapter therefore calls for a more critical approach to the selection and implementation of methodologies and approaches, and argues that this may be more important than the actual method selected. The authors hope that a greater awareness of the diversity of methods available to these researchers form an important part of this.

Chapter 7, *Scanning Horizons in Research, Policy and Practice*, focuses on the kind of methodologies that Chapter 6 argues may be most under-utilised, relative to its value, within Existential Risk Studies: structured expert elicitation. In particular, this chapter surveys different approaches to horizon scanning: collecting information from a wide range of sources and then using communities of practice to sort, verify, and analyse that information to look for early indications of poorly recognized threats and opportunities. This does not necessarily involve the quantification of estimates of risk, or any part of it, but does involve some kind of collective judgement about the relevance, importance, and or urgency of future threats. There are two forms of horizon scanning, both of which are relevant to Existential Risk Studies. The first is the exploratory — to identify novel issues by searching for the first signals of their emergence — and the second is "issue-centred", where we monitor issues that have already been identified as potentially emerging, confirm or deny this, and provide further judgements about them. Drawing primarily on the experience of the conservation community (which has played a significant role in developing these tools), the chapter assesses a range of techniques that can be used for these purposes and their implementations. It covers both manual and semi-automated approaches with regard to scope selection, input gathering, data sorting, cataloguing and clustering, result analysis and prioritisation, output utilisation, and process evaluation. The authors find that manual approaches require a structured form of expert elicitation to mitigate biases and promote objectivity but that semi-automated tools and AI may increasingly enable searches to be more impartial. For policy purposes, it is best if horizon scanning is actively incorporated by organisations and decision-makers into the policy design process, or at least if additional tools like road mapping are used to translate findings into policy, rather than when horizon scanning is used as a predominantly academic tool. Within CSER these findings have informed our own developing use of horizon scans, which we have predominantly applied to a range issues around Global Catastrophic Biological Risk. For instance, we continue to develop our horizon-scanning protocols and to seek out policy partners to implement them with, such as the World Health Organization.[2]

The next chapter, *Exploring Artificial Intelligence Futures*, surveys a different set of methodologies in relation to a different risk area, looking at methods and approaches for exploring possible futures for artificial intelligence that are accessible to researchers from the humanities. While they do not predict the future of AI, or its impact on society, these methods can still help us expand the range of possible futures we consider, to reduce unexpected surprises,

[2] World Health Organization. *Emerging Technologies and Dual-Use Concerns: A Horizon Scan for Global Public Health* (2022); World Health Organization. *Emerging Trends and Technologies: A Horizon Scan for Global Public Health* (2022).

and enable constructive conversations about the kinds of futures we would like to produce or avoid. Historically, one of the most influential approaches for thinking about the future of AI (and other technologies) has been the construction of speculative fictional narratives, as discussed in Chapter 1. While many of these have little value in thinking about AI, and are more focused in either using it as a metaphor for other issues or merely as an exotic narrative device, some have shown careful research and consideration and have had a significant role in provoking, and informing, public and professional discourses. However, fictional narratives tend to suffer from a range of issues, many of which stem from the economic incentives of the creative industries that produce them; such as the need to entertain audiences, pressure to embody AI in physical forms, like robots, that are more easily described or pictured, a lack of diversity in authorship and representation, and limited accountability for their claims. There is also a long history of researchers from a variety of disciplines, including Science and Technology Studies, philosophy, engineering, and risk analysis, producing high quality studies on the future of AI that draw on the unique insights from their field. Their predictions tend to fare poorly due to biases, partial perspectives, non-linear trends, and hidden feedback mechanisms. Furthermore, inevitable disagreements between experts across disciplines can have a paralysing effect for audiences. Group-based future explorations can address some of these challenges, using techniques like expert surveys, polling, interdisciplinary futures exercises, and expert elicitation, and there are also opportunities to extrapolate futures from past and current data trends. Perhaps the most promising methodologies are participatory futuring tools such as workshops, scenarios, and role-plays. These not only provide opportunities to combine divergent exploration with individual exploitation of knowledge and experience (as discussed in Chapter 5) but also offer greater accountability and opportunities for public participation, to promote responsible research and innovation (an important consideration also discussed in Chapter 2). However, they still need to be realistic, integrative, and data-driven, and producing such scenarios is a challenge that CSER has sought to address in later projects such as our participatory scenario role-play tool Intelligence Rising.[3]

Having surveyed a wide range of methodologies and approaches to studying extreme global risks, the second set of three chapters in this section focus on the development of three specific tools that CSER use to help us with our work.

Chapter 9, *Accumulating Evidence Using Crowdsourcing and Machine Learning: A Living Bibliography About Existential Risk and Global Catastrophic Risk*, describes the creation of a semi-automated process for systematically reviewing the relevance of academic research to the study of existential risk to improve our evidence base for policy and risk analysis. As Chapter 7 describes, this move to semi-automated scanning can be helpful for reducing individual biases and increasing the scope of research. In a systematic review, one of many time-consuming tasks is to read the titles and abstracts of research publications, to see if they meet the inclusion criteria. This chapter shows how this task can be shared between multiple people (using crowdsourcing) and partially automated (using machine learning). The authors used these methods to create The Existential Risk Research Assessment (TERRA), which is a living bibliography of relevant publications that gets updated each month and is freely available at terra.cser.ac.uk. The chapter presents the results from the first 10 months of TERRA, highlighting the potential and challenges of this tool and recommending that, for now, such semi-automated tools should only be used in tandem with manually curated bibliographies. The challenges noted include the need to make trade-offs between recall (inclusion of all relevant research) and accuracy (exclusion of irrelevant research), conflicts and inconsistencies in the assessment of papers for use in training the algorithms, and the incomplete assessment of this training data.

Chapter 10, *The Mortality of States Dataset*, describes the creation of MOROS (the MORtality Of States) database. This combines data on the lifespan of political states into a tool to help us understand both the phenomenon of collapse and the nature of entities that dominate global

[3] Avin, S., R. Gruetzemacher and J. Fox. 'Exploring AI futures through role play', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020): 8–14.

risk. The database defines a state as "a set of centralized institutions that coercively extract resources from, and impose rules on, a territorially circumscribed population" and characterises their lifespan in relation to "rough, critical dates in which significant changes to state form, function, and/or sovereignty occurred". The database was synthesised from a variety of primary data sources verified and expanded with a wider literature review. However, creating this tool was not simply a matter of data collection; significant interpretation was required to conceptualise states and their lifespans, and the authors ultimately view the result as a qualitative overview of expert opinion. In future it is hoped to use expert elicitation and structured literature reviews to improve the database, alongside finding better ways to code for the continuity of states and adding details about the consequences and reasons for state termination.

Finally, Chapter 11, *ParEvo: Enabling the Participatory Exploration of Alternative Futures*, describes ParEvo, an online tool for developing, and evaluating, alternative future scenarios using a participatory evolutionary process. ParEvo was developed by Rick Davies and has been applied by a range of organizations for different purposes. Since 2021, CSER has run three ParEvo exercises, two of which were used to inform this chapter. The ParEvo process is designed to be used by multiple people to produce both a set of storylines and data on the structure of participation in how people have collaborated to produce those storylines. ParEvo can help its users both to think creatively about alternative futures and about how they do that thinking while also prompting participants to consider ways of responding to possible futures and to exploit and/or mitigate their consequences. In general, ParEvo is more about the exploration of divergent futures, but as this chapter discusses, it could also be used for the purpose of convergence and exploitation of collective judgement. Due to its dual aims, ParEvo is not merely a scenario-generating tool but is also deeply tied up with the evaluation of scenarios and participation. The chapter considers the challenge in evaluating storylines, including establishing two-way feedback between participants and exercise facilitators; assessing post exercise outcomes; and identifying and assessing meta-goals for ParEvo as a tool.

All of the methods, tools, and approaches described above are presented within a particular context and often these relate to specific drivers of risk. Horizon scanning is discussed mainly in relation to conservation, due to the unique role that the conservation community played in developing these tools and bringing them to CSER. We have developed a scenario role-play tool, Intelligence Rising, only thus far used for exploring risks around Artificial Intelligence. MOROS shows the potential for historical data to be used to improve our understanding of collapse and the role of states. TERRA has so far only been implemented to assess literature concerning existential risk as an entire phenomenon, while the chapter on ParEvo considers two exercises that used it to explore possible futures for biotechnology governance and the field of Existential Risk Studies. As Chapter 6 argues, however, one beneficial development within Existential Risk Studies could well be the application of methodologies that have predominantly been used by one discipline or community to study a wider range of phenomena.

A lack of familiarity with different methods, tools, and approaches carries the risk of endowing them with a sense of "magic" to the uninitiated. It seems relatively common within academia for people to evaluate the usefulness of a particular method relative to their own familiarity with it, although it may be a matter of personal temperament whether people assume that methods they do not themselves use are "clearly useless" or "clearly beyond my understanding and therefore powerful". We hope that greater familiarity with different methodologies may dispel some of these biases and help researchers to appreciate what different methods, tools, and approaches have to add to our understanding of risk, and what they merely repackage into another intellectual form. This applies both to the methodologies we have described in detail here, like horizon scanning, scenario role-plays, and semi-automated literature reviews, and to those that we have not, like models, Bayesian reasoning, and analytical frameworks.

In the next section we will turn to considering what these models are generally used for: the assessment and management of different causes of risk. Some of the chapters of that section clearly utilise one of these methodologies and some draw on several. We hope that having familiarised themselves with the chapters in this section, however, readers will be in a better position to both understand how these methodologies are being used and what they can really tell us about different risks and to critically reflect on the quality of their implementation and the resulting conclusions that are drawn from this.

6. An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards

SJ Beard, Thomas Rowe and James Fox

Highlights:

- This chapter examines and evaluates the range of methods that have been used to make quantified claims about the likelihood of existential hazards, drawing on a comprehensive literature review of 67 such claims, across 13 kinds of hazard, that are presented in an appendix.
- The chapter uses an informal evaluative framework to consider the relative merits of these methods, regarding their rigour, ability to handle uncertainty, accessibility for researchers with limited resources, and utility for communication and policy purposes.
- The authors find that each method has advantages and disadvantages but there is no uniquely best way to quantify existential risk. Nevertheless, they argue that some methods may be more suitable to certain purposes within Existential Risk Studies and that some methods, especially the Delphi technique and related forms of Structured Expert Solicitation, should be more widely used within the community.

https://doi.org/10.11647/OBP.0360.06

- More importantly, however, they find that, in many cases, claims based on poor implementations of these methods are still frequently invoked by the Existential Risk Studies community, despite the existence of more methodologically robust estimates.
- The chapter ends with a call for a more critical approach to the selection and implementation of methodologies and the use of quantified claims within Existential Risk Studies, and argues that a greater awareness of the diverse methods available to these researchers should form an important part of this.

This chapter was published in 2020 after a multi-year review of evidence and methodology by the authors and a preliminary presentation and discussion at the 2018 Cambridge Conference on Catastrophic Risk. Its calls for more critical reflection on methodological selection and implementation, as well as for the greater use of structured expert solicitation, continue to inform the developing field of Existential Risk Studies. The uses of horizon-scanning tools like Delphi is further discussed in Chapter 7, while an applied example of this type of methodology is continued in Chapter 15. Chapter 10 meanwhile offers an example of using historical data in Existential Risk Studies and Chapter 13 shows one method for constructing system models from empirical data.

> Apocalyptic predictions require, to be taken seriously, higher standards of evidence than do assertions on other matters where the stakes are not as great.
>
> — Carl Sagan[1]

How likely is it the next century will see the collapse of civilisation or the extinction of humanity, and how much should we worry about different hazards that constitute this risk? These questions seem to bedevil the study of existential risk. On the one hand, they are important questions that deserve answers, not only to assess the level of risk facing humanity but as part of an integrated assessment of existential risk and opportunities to mitigate it.[2] On the other hand, the quantification of existential risk is extremely challenging. As Carl Sagan pointed out, part of the problem with apocalyptic pronouncements is their theoretical basis and the fact that they “are not amenable to experimental verification — at least not more than once".[3] Not only would a human extinction event be unprecedented, but the risk of such an event tends to emerge at the interaction of complex social, environmental and economic systems that are hard to model, and there is a substantial degree of uncertainty about the second-order "risk space" where they might be found. Together, these problems have made the quantification of risk speculative and reliant on new and creative methods for analysing the threats humanity faces.[4]

Previous analysis by Bruce Tonn and Dorian Stiefel attempted to resolve this tension, by evaluating a range of methods for quantifying existential risk from an ideal perspective, and made several important recommendations, most of which we endorse.[5] However, although the number of researchers developing methods for the quantification of existential risk has grown, these methods have only been applied in a piecemeal fashion to a limited number of disciplines. For instance, conservation biologists have made important innovations in the use of structured expert elicitation; analysts in the Intelligence Community have developed new forecasting techniques with a high degree of success in assessing the probability of events in the near future and climate scientists, economists and epidemiologists have developed powerful models of complex global systems. These, together with the range of existing techniques — from philosophical analysis and opinion polling to toy modelling and fault trees — each have benefits and limitations, and we believe it is time they were more widely understood and adopted. The aim of this chapter is, therefore, to survey the literature on the quantification of existential risk, to introduce different techniques to new audiences and to give an informal assessment of their capabilities, together with some suggestions for how they can be implemented and improved. We hope that this will spur still more methodological diversification and development.

This chapter is addressed primarily to a group of scholars we refer to as the Existential Risk Studies community. This is made up both of researchers who work within institutes that focus on the study of existential and Global Catastrophic Risk (such as the Future of Humanity Institute, the Centre for the Study of Existential Risk, The Global Catastrophic Risk Institute and the Future of Life Institute) and those who are consciously seeking to align their research with the goal of understanding and managing such risks. However, it is written from an awareness that these two groups do not necessarily contain all of those researchers whose work can be expected to contribute to our understanding and management of existential risk, and indeed many of the sources we consider were produced by researchers who are in neither of these groups. It is our conviction that one of the key roles of existential risk organisations like the above should not only be to support researchers who see themselves as falling into the Existential Risk Studies community but to forge better connections with, and a better understanding of, all research that is relevant to the understanding and management of these risks.

Finally, it is worth noting that this chapter does not seek to consider fully every aspect of the quantification of existential risk. The risk arising from a specific threat is given by multiplying the probability of the threat occurring with its expected severity. This chapter only sets out to examine the methodologies used to quantify the former. If one held the severity of all threats constant, at the point of human extinction, then this is all one needs to know. However, while some studies in this chapter do aim to assess this, others assess threats at a lower point of severity, such as that of causing a global catastrophe or civilisational collapse, while others do not specifically consider the severity of a threat but are included rather because they relate to potential scenarios that have been of interest to scholars of existential risk. In these cases, one needs to be mindful that the severity of the event is still to be determined, or at best is only imprecisely defined when considering the overall quantification of that risk.

1.1 A brief introduction to different notions of probability

We begin by noting that, while this chapter refers to the probability of existential risk, there are multiple ways of understanding the notion of probability. The first of these is the frequentist, or objective, notion of probability. According to this approach, probabilities are fundamentally related to the frequencies of events based on past observations. Once an experiment has been repeated many times, the frequency of any observed phenomena indicates its underlying regularity. Therefore, frequentist probability claims are sensitive to the experimental setup and measuring technique, and any new evidence requires probabilities to be reassessed from scratch. The second notion of probability is the Bayesian, or subjective, account, according to which probabilities represent our level of belief that a phenomenon will occur. One begins with a subjective prior belief about the probability of an event and then updates this via Bayes' Theorem (or Bayes' rule), which specifies how additional information affects the probability of an event. Even though these probabilities are subjective, one is required to set out reasons for arriving at them, which allows others to challenge these or update them further.

In this chapter, we discuss the two notions of probability interchangeably with only a few passing remarks and we will instead focus on the quality of the methodologies that can be used to produce both kinds of probabilistic statement without prejudice. This is because even though Bayesian notions of probability dominate the field of existential risk, there are some areas — most notably in the domain of public health — where frequentist notions of probability are more common.

1.2 Conceptual challenges in studying existential risk

The term "existential risk" can be understood in many ways, and clarifying its definition is undoubtedly a crucial concern. However, since our aim is merely to study methods of quantifying existential risk, which approach this term from multiple angles, we will take a very broad view of how the term should be used, encompassing human extinction, civilisational collapse and any major catastrophe commonly associated with these things.

Another point is that most studies consider existential risk in terms of distinct threats (such as nuclear war, pandemics and climate change). However, global catastrophes tend to involve a combination of multiple factors, including a precipitating catastrophic event, a systemic collapse that spreads this catastrophe to the global scale and a failure to take adequate steps to mitigate this risk.[6]

Finally, existential risk cannot be studied in a vacuum. Even our assessment of such risks can profoundly affect them. For instance, if we take existential threats more seriously, this may lead to greater efforts to mitigate them. Sometimes, risk assessments can take account of human activities, such as when multiple estimates of catastrophic climate change reflect different future emission paths; however, this is not always possible. Whilst important, we do not see this as a problem that we must solve here, since it is common to many fields of risk analysis and affects all the methods we describe to a greater or lesser extent.

Whilst some futurists may respond to these difficulties by adopting a pluralist conception of multiple futures, in which the goal is to map out the likely consequences of decisions that we face in the present, existential risk mitigation must go beyond this. In particular, it is necessary for research and mitigation efforts to be prioritised and for risk-risk trade-offs to be undertaken, such as when assessing dual-use technologies; these require the quantification of risk. We therefore believe that it is imperative to combine such pluralistic future scenarios into an integrated assessment that takes account of factors such as the resilience of global systems and the magnitude of existential risk. Given that such assessments are at an early stage, however, we will generally assume that all risk assessments are being made against a "business as usual" scenario, where people continue mitigating risk roughly as much as they did at the time when that risk was assessed. In general, we suspect that this will overstate the future level of risk because it misses the potential for technological and governance interventions. However, that may not always be the case, as economic development can systematically push global systems into a more fragile state, making existential risk increase over time.

1.3 Four criteria for evaluating methodologies

As well as presenting and discussing the existing methods for quantifying existential risk, we will provide an informal assessment of each according to the following four criteria:

Rigour: Can they make good use of the — generally limited — available evidence? Three key considerations for this are: 1) their ability to access a broad range of information and expertise from across multiple perspectives, 2) the suitability of their means for turning this into a final judgement, and 3) the ease of incorporating new information into this judgement or combining different judgements together using the same method.

Uncertainty: How well do they handle the — generally considerable — uncertainty in this field? Three key considerations for this are: 1) whether they provide opportunities to quantify the level of confidence or uncertainty in their estimates, 2) whether the application of this method tends to systematically ignore or compound sources of uncertainty in the process of forming a final judgement and 3) whether they can help to identify and overcome epistemic bias.

Accessibility: Can they be applied by the individuals and — generally small and interdisciplinary — research groups that make up the Existential Risk Studies community? This is to be assessed in terms of: 1) the amount of time required to implement them in a reasonable way, 2) the level of expertise required for a researcher to take a lead or principal role in implementing them and 3) what other barriers exist to their implementation.

Utility: Do they provide results that can be used for purposes like policy selection and prioritisation, and can be communicated to varied stakeholders? Three key considerations for this are: 1) their credibility, both with scientists and non-scientists, 2) their ability to provide useful quantified information and 3) their ability to provide further information and insights about a risk and how to manage it.

For each of these, we assess methods on a four-point scale from Very Low to High and summarise the results of this evaluation in Appendix B. We do not mean to imply that these four criteria are of equal importance; however, their relative importance is likely to depend upon the context for which an assessment of existential risk is being produced.

2. Analytical Approaches

Not all methods for attempting to quantify existential risk are based on specific evidence for that risk. Given the lack of evidence available, this is perhaps more appropriate than it seems.

By far the most widely discussed of these is the so-called "doomsday argument", which was developed by multiple philosophers including Brandon Carter, Richard Gott and John Leslie. It is a statistical argument about the probability that any given human observer will be at a particular place in human history. There are multiple versions of this argument. Some appeal to our current position in the timeline of human history — for instance, it is much more likely that human history thus far represents more than 5% of all human history than that we have more than 95% of our history ahead of us. Others appeal to our position in the population of all human observers — it is much more likely that the human population born before me represents at least 5% of all humans than that 95% of humans will be born after me.

These arguments are controversial and they have been used to justify very different claims about the probability of existential risk, from a less than 2.5% chance that we will fail to survive for at least another 5,100 years,[7] to a 20% probability that we will not survive the next century.[8] Important factors that determine the difference between these claims include the type of doomsday argument being deployed, one's assessment of the length of human history or the number of humans who will ever exist and one's assumptions about the temporal distribution of existential risk (whether it is evenly distributed over time or comes in peaks and troughs). In theory, some of the uncertainty about these could be quantified by providing multiple calculations with differing assumptions; however, in practice, this is seldom done.

Two other analytical arguments are sometimes discussed in relation to their implications for the likelihood of human extinction. The first is known as the Simulation Argument, which connects the probability that humans face imminent extinction to the probability that we are living in a computer simulation.[9] The second, known as the Great Filter Argument, connects the probability that humans face imminent extinction to the probability that there is intelligent life on other planets.[10] Several sources cite a 2006 working paper titled 'The Fermi Paradox: Three Models', by Robert Pisani of the Department of Statistics at UC Berkeley, as providing a quantification of existential risk based on the Great Filter argument. However, the paper was never published, and no version is currently available online. Apart from this, these arguments have yet to be used for the purpose of existential risk assessment.[11]

Part of the explanation for the interest in, and use of, the doomsday argument and other analytical tools is their high level of accessibility. They score Very Low across the other three categories. However, this does not mean these approaches should have no role to play in quantifying risk, as they can still inform people's prior beliefs about human extinction, i.e. what we might most reasonably guess about the likelihood of human extinction before considering the evidence. A Bayesian account of probability requires such a prior to begin. In most scientific cases, the prior's precision is relatively unimportant as one can continually update it with evidence. However, within the context of existential risk, where evidence is sparse, prior probabilities are disproportionately important, meaning that any technique helping produce better priors can still be very useful. This is the approach taken by Leslie[12] and Wells.[13]

3. Modelling-Based Approaches

Since it is not possible to undertake an empirical study of human extinction or global civilisational collapse, the next set of methods use observable evidence to produce a set of assumptions, or a model, that allows us to study them indirectly.

3.1 Toy models and extrapolation from data

The simplest approach of this kind involves assuming there exists an underlying regularity in the frequency of certain events that have historically already occurred and have the potential to pose an existential risk in the future. A frequentist analysis of historical data can then be used to estimate the approximate time interval between such events and hence produce an annual occurrence probability. So far, this approach has been applied to asteroid impacts,[14] supervolcanic eruptions,[15] nuclear wars,[16] space weather,[17] particle physics experiments,[18] and the occurrence of extinctions and global catastrophes in general.[19] Yampolskiy also applies this approach to predicting AI failures but without producing a quantified estimate of their probability.[20]

Where no such underlying frequency can be assumed — for instance, because an event is historically unprecedented — it is possible to produce simple toy models that can allow one to use historical data to determine the probability of an event occurring in other ways. Firstly, one can assume that the magnitude of impact from certain threats follows a specific distribution, enabling one to estimate the probability of a large impact event taking place from the historical record of smaller impact events.[21] For instance, Millett and Snyder-Beattie[22] for the number of fatalities from biowarfare or bioterrorism and Riley for solar flares both assume power law distributions,[23] while Bagus assumes that the fatalities from influenza pandemics follow an exponential distribution.[24] From this, they can estimate the probability of more extreme events of this kind that have the potential to pose and existential risk. Other toy models include Day, André & Park,[25] Fan, Jamison and Summers,[26] and Millet and Snyder-Beattie,[27] all of which assess the risks of catastrophic pandemics.

Secondly, one could assume that a currently unprecedented event would occur as a consequence of multiple events with historical precedents. The existential event could be the end result of a chain of precipitating events each a possible consequence of another. In this case the unprecedented event's probability is the product of the conditional probabilities down the chain. Klotz and Sylvester,[28] Lipstich and Inglseby,[29] and Fouchier[30] use this method to estimate the likelihood and impact of a global pandemic resulting from Gain of Function influenza research by assuming that such a risk results from the occurrence of two events whose probabilities are easier to determine, a laboratory-acquired infection and a biosecurity failure. Similarly, Hellman estimates the probability of a nuclear war resulting from a "Cuban Missile Type Crisis" as the product of probabilities of a sequence of four precipitating events.[31] Alternatively, the unprecedented event could arise due to the coincidence of two mutually independent events. This would create the basis for a fault tree, discussed in the next section.

Although these approaches are methodologically clear, clarity does not imply objectivity. For instance, the analysis of the geological record is itself speculative and open to interpretation,[32] so that Decker's estimate of the probability of supervolcanic eruptions,[33] while widely cited in the Existential Risk Studies community, is often seen as pessimistic amongst volcanologists and is higher than most other estimates. Other instances of disagreement are even greater, Lipsitch and Inglesby[34] and Fouchier[35] arrive at probability estimates over seven orders of magnitude apart, despite using the same toy model to evaluate the same risk. This points to the central issue, approaching the same historical data in different ways can result in very different probability estimates. Love comments on the "limited accuracy of statistical estimates" when comparing his result to Riley's for the recurrence probability of Carrington-like geomagnetic storms, saying that we can only conclude that the probability is "somewhere between vanishingly unlikely and surprisingly likely".[36] Meanwhile, Hellman openly cherry picks evidence so as to not appear "alarmist".[37]

Another problem with these methods is that some events are excluded from the historical record because of the "anthropic shadow" they would leave. Roughly speaking, were any such event to have occurred in the past, this would have led to the non-existence of the present observer. Therefore, even if its probability was extremely high, it must seem to us as if it could not have happened because our very existence depends on it.[38] David Manheim looked at risk estimates of natural pandemics and concluded that there is significant uncertainty about the relationship between historical patterns and present risk because of such "anthropic factors and other observational selection biases".[39] Tegmark and Bostrom also take account of this effect when quantifying the threat from particle physics experiments, but most theories ignore it. A related issue is that the historical record may need to be understood not just by what happened but also what didn't happen. Gordon Woo argues we should sometimes incorporate "counterfactual analysis" of near-miss events to more accurately model risks because of the likelihood that very rare events will be underrepresented in historical data.[40] For instance, if the true underlying probability of a certain sized asteroid striking earth was 0.004 per year and we have 100 years of data, then it is most probable that this event would not have occurred within this period leading our analysis to underestimate its probability. However, if we had evidence of asteroids of this size passing close to the earth without striking it during this period then we can use this information to arrive at a better estimate for the underlying probability of such a strike. Woo cites numerous "near-misses" in fields from maritime and air disasters to terrorism that can be used to make better predictions about rare kinds of catastrophe,[41] while other scholars have incorporated historical near misses into the study of nuclear war.[42]

In principle, using historical data allows us to calculate our degree of uncertainty via simple statistical techniques. Thus, many of the sources that utilise this approach provide ranges for the probability of the events that they study. However, care needs to be taken when combining uncertain statistics in toy models, since merely multiplying uncertainty ranges will overstate uncertainty as the coincidence of multiple outliers can be expected to be rarer than any individual outlier. Furthermore, because the modelling approaches we discuss make assumptions about the distribution of the underlying phenomena from which the data has been sampled, they are vulnerable to abnormal (or even discontinuous) changes in these phenomena. For instance, Fouchier points out that when using historical data about biosecurity, one must account for generally increasing safety standards. Similarly, Manhiem notes that the likelihood of Global Catastrophic Biological Risks might be higher than historical evidence implies because of contemporary global travel, high population densities in megacities, and closer contact with animal populations due to factory-farms.[43] Therefore, simplistic frequentist approaches may underestimate the appropriate level of uncertainty.

Despite its simplicity and popularity, we believe that toy modelling of existential risk has serious shortcomings. We rate these methods Low for rigour, uncertainty and utility, while once again noting their often underutilised potential to quantify and estimate uncertainty. Indeed, the utility of this approach appears to be highest in cases where assessors are overconfident, both regarding the uncertainty surrounding their predictions and the objectivity of the historical record, which is potentially dangerous. We do however rate this method highly in terms of its accessibility and the opportunities it provides for researchers to bring existing evidence to the study of existential risk. We see this approach as having limited appeal for quantifying existential risk as the field matures, though it may still play a useful role in stimulating further research and in producing estimates that can be taken as "objective" — priors for further Bayesian analysis.

3.2 Fault trees and Bayesian networks

A more sophisticated modelling technique for studying existential risk involves fault trees. Originally developed to model the emergence of system failures in safety engineering, they have now been widely applied in risk analysis. Fault tree models use Boolean algebra to map out in a logic tree, using "and" and "or" gates, how a system failure could arise. Branching backwards from overall system failure at the tree's top, we first write the ways failure could happen as different nodes and then branch further backwards with how this node could fail and so on. If possible, a probability of failure is assigned to each node and then one can sum or multiply probabilities, depending on the Boolean nature of each gate, to give the overall probability of system failure. Importantly, fault trees can also clearly reveal preventative steps that could be taken. This technique has been used to study risks from nuclear war[44] and AI safety.[45]

Bayesian networks are an extension of fault trees and present great promise for existential risk analysis.[46] Although the power of Bayesian networks is also rooted in its graphical representation of the possible failure under study, the nodes now represent random variables, and it is the edges between these nodes that are quantified. These edges are directed from parent to child nodes, and every node comes with a conditional probability table to provide the causal probabilistic strengths for the edges between connected nodes. As with fault trees, a Bayesian network is first drawn by working backwards from a failure state (or any other outcome one wishes to study) through the conditions that might lead to this state. However, unlike fault trees, Bayesian networks can handle dependencies between different parts of the system and so all conditions can be factored into one's analysis, including those that are only rarely important. This can make it easier to incorporate information from near-misses and other tenuous sources whilst still being rooted in observed system behaviour. Once a network has been created, Bayes' rule can determine the final expected occurrence probability of each node in this network, including that of the final outcome, and this can be calculated dynamically and updated continuously to take account of additional information.

Khakzad et al. show that fault trees are less suitable for modelling complex systems because Bayesian networks are far superior at handling dependencies between different parts of the system, common cause failures, and uncertainty.[47] Also, Bobbio et al. demonstrated that every fault tree has a corresponding Bayesian network and so the methodology is proven to be generalisable.[48] However, when evidence is sparse, the results from Bayesian networks can be significantly affected by the modeller's choice of prior probabilities. Bayesian network analysis has not yet been used to study existential risk directly; however, it has been applied to a variety of catastrophe models — for instance, Li et al. demonstrated how Bayesian networks can assess catastrophic risks under uncertainty by modelling catastrophic flooding in China.[49] Both fault trees and Bayesian networks improve on simpler toy models as they can manage more complex system dynamics whilst handling a greater range of data and inputs. We rate both approaches as Medium for utility but note their particular value in providing insights that can be used to study risk mitigation by modelling how changes in components of a system affect its probability of failure. To be useful, the model must be a sufficiently faithful and detailed representation to capture accurately the effect of individual policies or interventions. It is possible that if the model fails on either of these, then it may simply lead decision-makers astray and give them false confidence. This is reflected in our higher degree of uncertainty about this particular classification.

Bayesian networks outperform fault trees in terms of both rigour (High/Medium compared to Medium) and uncertainty (High compared to Low), performing especially well in relation to the quantification of uncertainty. However, fault trees outperform Bayesian networks in terms of accessibility (Medium compared to Low), a category in which neither method performs well as both require significant modelling skills and domain knowledge. Fault tree analysis requires considerable familiarity with the underlying system to understand the processes that may lead to it failing. Bayesian networks require an even greater degree of familiarity as the assessor must provide probabilities for all the conditional relationships that may play a role in determining an outcome.

3.3 Adapting and applying existing models

Some researchers, usually from outside the Existential Risk Studies community, have also adapted existing models to study globally catastrophic and existential threats. These include using models of pandemic influenza to assess the likelihood and impact of a "modern Spanish flu",[50] adapting IPCC climate models to assess the probability of catastrophic climate change[51] and using astronomical models of near-Earth objects to assess the likelihood of asteroid impacts.[52] Other authors infer from existing estimates of the uncertainty in models what the possibility of more severe impacts might be, including: Wagner and Weitzmann for catastrophic climate change[53] and Atkinson et al. for near asteroid impacts.[54] Finally, some identify parameters where they believe the model is mistaken and use this as a justification to "correct" the final output, as Dunlop and Spratt[55] and King et al. do for the IPCC's predictions about climate change.[56]

In theory, these approaches can build on the underlying rigour and utility of the model being used; however, this depends significantly on the validity of their adaptation or application. Most models, including the IPCC's, are not designed to assess catastrophic risks specifically, so catastrophic outcomes will be outliers. This provokes debate about whether these outcomes should be treated with genuine concern or can be dismissed as model failures.[57] Modifications must only be made by those with significant skills and a deep understanding of the model's functionality to adapt them in ways that will preserve their virtues. This may explain why the most comprehensive efforts to do this come from scholars like Ramanathan and Madhav who are outside of the Existential Risk Studies community, which, in turn, may mean that these researchers understand less about the nature of such risk.

Despite their very low levels of accessibility, well-executed applications of sophisticated modelling techniques represent a desirable next step in the study of Existential Risk and we rate them highly in terms of rigour, uncertainty and utility. However, we are sceptical of attempts to replicate the success of high profile existing models with fewer resources, by making less considered model adjustments or making concrete predictions based solely upon their current levels of uncertainty and suggest that this approach should be adopted cautiously and only by better resourced groups within the Existential Risk Studies community.

4. Subjective Approaches

Of the 66 sources in our literature review, 45% relied, at least in part, on the subjective opinion of the author or others, without direct reference to specific models, data or analytical arguments. This included all the sources that discussed the potential threat from Artificial Intelligence, which many existential risk scholars believe to be the most significant. This is unsurprising given the difficulties that other methods face, and the use of subjective expert opinion is a well-established and successful means for handling uncertainty in many other scientific fields.[58] However, not all subjective opinion should be treated equally and, in the next two sections, we will consider different approaches.

4.1 Individual judgements

At one end of the spectrum are individual opinions given without reference to clear reasoning. Examples include Rees[59] and Stern,[60] who both consider the overall probability of human extinction, one for a "scientists' warning" and the other to determine the correct social discount rate. Bostrom provides a similar judgement, although this appears to be based on updating a prior belief derived from analytical arguments, to account for "the balance of evidence".[61] Halstead[62] and Chapman[63] both present a considerable degree of evidence and argumentation before offering subjective conclusions about the threat of climate change and asteroid impacts, but without any specific method to connect the two. At best these estimates represent what Tonn and Stiefal refer to as "holistic probability assessments", in which "the individual probability assessor estimates the holistic extinction risk through informed reflection and contemplation".[64]

A more sophisticated approach to subjective opinion formation is for probability assessors to break down any risk into a set of mutually exclusive threats and then classify the danger posed by each of these and the probability of their occurrence (together with the likelihood that they would pose an existential risk). Tonn and Stiefel refer to this as the "whole evidence Bayesian approach". This encourages a systematic way of estimating probabilities and is useful to anyone reviewing such an assessment as it makes it easy to update predictions in the light of new evidence or different reasoning.[65]

Another approach is what Tonn and Stiefel refer to as "Evidential Reasoning". This involves specifying the effect every piece of evidence has on one's beliefs about the survival of humanity. Importantly, these probabilities should only reflect the change that this evidence makes, not one's initial prior beliefs, allowing others to assess them independently. As such, they will only be "imprecise probabilities" that describe a small portion of the overall probability space, where the contribution each piece of evidence makes to one's belief and its complement need not sum to 1. For instance, one might reason that evidence about the adaptability of humans to environmental changes suggests a 30% probability that we will survive the next 1,000 years, but only a 10% probability that we will not.[66] Combination functions can then be used to aggregate these imprecise probabilities to return the overall probability of extinction within this period. This method not only helps assessors determine the probability of extinction, but also provides others with information about the sources of evidence that contributed to this decision and the opportunity to determine how additional information might affect this.

A final method listed by Tonn and Stiefel draws on the technique, common amongst futurists, of anchoring assessments in scenario-based considerations of what it would take to bring humanity to extinction. Assessors envisage a possible human extinction scenario and then consider how indicative this scenario is of both the space of all possible future scenarios and the space of those in which humanity goes extinct. This exercise is repeated until the assessor judges that they have exhausted all, or at least a substantial portion, of the human extinction scenario space. They can then estimate what proportion of future scenarios involve human extinction, and by extension how likely this is. Advantageously, this focuses on the end result, human extinction, rather than on the processes by which this might be brought about, although the use of scenarios is sometimes frowned upon in other communities. Despite the fact that these three techniques consist of little more than clearly setting out one's assumptions and reasoning process for others to follow, none of them has so far been implemented well in the literature on Existential Risk.

It has been shown that, with only a few hours of basic level training using freely available tools, most people can be calibrated to give reliable estimates of their level of uncertainty for their subjective opinions based on their current state of knowledge.[67] Despite this, few who have conducted subjective probability assessments have indicated that they have undertaken any such calibration or to state their degree of uncertainty. Instead, experts have tended to hedge their bets merely by couching otherwise precise estimates in vague language. Furthermore, individuals routinely suffer from overconfidence and confirmation bias in their subjective estimates, and when individuals have their name attached to a figure, such biases become especially problematic. Eliezer Yudkowksy discusses the relevance of cognitive biases affecting the judgement of Global Catastrophic Risk including the availability heuristic, hindsight bias, conjunction fallacy, scope neglect and overconfidence.[68]

The popularity of individual subjective opinion is probably because they are especially easy for researchers to apply and are often offered as the basis for further discussion and inquiry in the future. Such estimates can also be well received by media and policymakers alike, especially when they can be linked to a high-profile academic of celebrity and hence become associated with that individual's perceived authority, potentially enhancing their utility. Indeed, it often appears easier to get people to agree with the single judgement of a known individual than a collective judgement which combines information from that individual with others.

The quality of individual subjective opinion thus depends on both the person providing the estimation and where suitable techniques are used to present and clarify their reasoning and assumptions. We rate this approach as Low/Medium for rigour. Despite the fact that this kind of estimate is well received we rate its utility as Low, reflecting its generally narrow focus and lack of credibility within scientific communities. We also rate it as Very Low for uncertainty, in particular due to its weaknesses in overcoming bias. However, we rate this method as high in terms of accessibility, which probably helps to explain its relative popularity in this field. The kind of robust approaches to clarify one's thinking would hardly detract from this high level of accessibility, and indeed may make it even easier for assessors to reach a final judgement, so it is disappointing to see them so little used.

4.2 Aggregating expert opinion

Another way of seeking to improve on individual subjective opinions is to pool together the judgements of multiple people to account for a more diverse range of perspectives. There are two reasons why this could improve the quality of judgements.

The first of these relates to the "wisdom of crowds"[69] which provides an epistemic justification for the aggregation of large numbers of individual opinions to determine the truth of some proposition. It relies upon the assumption that individuals receive some kind of signal pointing to the truth or falsity of that proposition, and that as a result, they are slightly more likely to judge correctly than incorrectly. The theory then states that, so long as individuals are making independent judgements, adding more will increase the probability that the group's median judgement will tend towards the correct one. The distribution of judgements across the group will effectively cancel out the noise that leads to some incorrect individual judgements and amplify the correct signal.

The second is that whilst individual judgements will be affected by multiple biases, when aggregated over many people, these biases may average out, improving collective judgement's accuracy. However, this can also be counterproductive as biases are often shared across large groups, or even reinforced by groupthink and the sense that one may be judged by biased peers. This violates the independence of individual judgements and can lead to the predictive power of a group decreasing with its size (Fujisaki et al., 2018). Partly as a response to this, some studies have suggested that aggregation methods that give more weight to outlying opinions outperform straightforward averaging approaches.[70]

In this section, we will limit ourselves to discussing approaches which simply average expert judgements, whilst in the next section, we will turn to more structured and deliberative approaches.

Simple aggregation is the dominant method for making predictions about the existential risk from Artificial Intelligence,[71] but has also been applied to the prediction of nuclear wars[72] and to quantifying existential risk in general.[73] These surveys vary considerably in quality and size, with many showing little concern for the diversity of participants, the statistical rigour of their analysis or uncertainty quantification (the honourable exception being Grace et al.).[74] Most surveys take the median response as their prediction, but Turchin argues that this is sometimes not optimal for existential risk.[75] For AI safety, instead of using the median estimate of AGI creation for risk assessments, we should be concentrating on the earliest possible time of AGI creation and define a "minimum acceptable level of AI risk".

Those who have adopted this approach often acknowledge its limitations. For example, Sandberg and Bostrom state that "these results should be taken with a grain of salt. [...] There are likely to be many cognitive biases that affect the result, such as unpacking bias and the availability heuristic as well as old-fashioned optimism and pessimism".[76] Judgement independence is hard to ensure as surveys are often completed at conferences and so it is difficult to guarantee that individual judgements are not influenced by others. Remote participation via an anonymous platform may offer a partial solution to this problem. However, given how close-knit many academic and technical communities are, this still may not secure judgement independence. Finally, although aggregation may improve judgements, it has the effect of making them less well behaved. For instance, if one seeks the median estimate from a group about the probability of superintelligence being developed and the probability of superintelligence leading to human extinction and then combines these figures, this can differ substantially from the median group prediction that humanity will go extinct from superintelligence.

The aggregation of expert opinion has the potential to improve upon individual judgements regarding their rigour and ability to handle uncertainty; however, in practice, this opportunity is, once again, often not taken. This may reflect the fact that extensive, well designed surveys are still out of reach of many small research groups and that people seem to respond equally well, if not better, to overconfident survey results from a small pool of "experts" than to extensive well-designed surveys that express an appropriate degree of uncertainty. We rate this approach as Medium for rigour, Low for uncertainty and utility, and Medium, both with a reduced level of confidence, for accessibility (due to disagreements about the amount of time it takes to conduct surveys in a "reasonable" way).

5. Structured and Deliberative Approaches

The final family of approaches we discuss also use subjective opinion, but seek to combine multiple opinions in more structured ways than simple aggregation. A variety of such methods have been developed by scientists and foresight specialists to aid decision-making under uncertainty, although so far these have been sparsely used in quantifying existential risk. Some of the techniques we describe form part of proprietary foresight tools such as the Delphi Technique (developed by RAND) and Superforecasting (developed by the Good Judgement Project); however, these can be disaggregated into their constituent parts for the purposes of discussion.

5.1 Weighted aggregation

The first of these approaches weights opinions differently in the aggregation based on an assessment of each individual opinion's value. For instance, Roger Cooke's "classical" approach to expert elicitation gives greater weight to subjective opinion based upon experts' performance on a series of calibration questions that ask them to predict things that are either known or that can easily be determined.[77] An expert who more often gets closer to the truth has a larger weight in the overall aggregation of judgments. This approach's prediction accuracy has been shown in multiple studies to outperform simple aggregation.[78]

However, this method is not well suited to predicting existential risk as the experts' competency at predicting catastrophic and existential risk cannot be calibrated due to their unprecedented nature. It might be possible to test experts' putative accuracy through their success at predicting more common and nearer term future events; someone's success at predicting short-term AI milestones could reflect the strength of their predictions about the long-term future of AI. However, there is no obvious means for assessing how success at predicting short-term and long-term trends are related.

An alternative means of weighing individual judgements is via peer ratings of respect and reliability. Theoretically, this avoids the problem of needing to calibrate individual predictors based on past performance and could help individuals assess their own beliefs by considering their credence in the beliefs of their peers. However, such weights are often of little, if any, meaning, especially in the context of existential risk.[79] Weightings can also be generated by repeatedly sampling experts' predictions and weighting those who gave more consistent answers more strongly than those whose answers varied, potentially indicating a lack of evidence-based judgement. For instance, Bamber and Aspinall asked for the same estimate from experts two years apart and those resampled were forbidden from referring to their first estimate to test the stability of individual judgements in determining the risk of a future sea level rise due to climate change.[80] However, this approach is problematic because it would be hard to distinguish between estimates that varied over time because of the randomness arising from a lack of evidence and those which changed because the estimators were successfully updating their predictions to take account of additional information.

The final method of weighting that we consider was developed by the Good Judgement Project on the basis that empirical studies indicate that some individuals (who the project terms "superforecasters") are substantially better at making predictions about the future than others. The project selected over 2,000 individuals and tasked them with assessing the likelihood of various world events. It found a considerable degree of variance amongst participants, with some individuals performing consistently well regardless of the kind of prediction they were being asked to make. Furthermore, it found that individuals who performed consistently highly in making accurate predictions were able to outperform even domain experts and professional intelligence operators. Philip Tetlock, the project's leader, concluded that these individuals had particular psychological traits that led them to make more accurate predictions, including caution about the strength of their beliefs, humility about the extent to which complex processes can be simplified, curiosity about the facts of a case, valuing diverse views and opinions and a belief in the possibility of self-improvement.[81] However, rather than assessing these psychological traits directly, the key to identifying superforecasters has been to keep track of individual performances at making predictions, including people's ability to update these in order to account for new information. This was done by assigning a "Brier score" to each superforecaster, an assessment of how close their predictions came to actual events.[82] The project found that the most accurate predictions were produced from an aggregation of participants' predictions, but those with the highest Brier scores were weighted more strongly. However, it is worth noting that the success of these superforecasters was found to diminish significantly when they were asked to make predictions more than 12 months ahead. At present it is unclear whether this reflects a limitation of superforecasters' abilities or a general problem with making longer term predictions.

While some evaluators rated these methods slightly more favourably our overall evaluation of them is no better than Aggregated Opinion Surveys for any category but low for accessibility. Given this, it is hardly surprising that such techniques have not been used in the assessment of existential risk thus far.

5.2 Enhanced solicitation

Another approach to structured expert elicitation is to seek to improve, rather than simply measure, the quality of experts' predictions. Broadly speaking, this can be performed prior to solicitations being made, at the point of solicitation or between solicitation and a final judgment being produced.

Pre-solicitation methods of improving the quality of expert judgement focus on training and method selection. We have mentioned a variety of such approaches already in Section 4.1, including calibration of uncertainty and the use of formal methods like evidential reasoning and holistic probability assessments. However, some methods specifically focus on prediction as a structured group activity and these are worth noting here. For instance, the Good Judgement Project found that both natural superforecasters and those who did not share their psychological traits were able to learn and develop them over time to greatly improve the accuracy of their predictions when they went through a process of probability training, teaming and tracking. Probability training helped correct cognitive biases, teaming allowed for the sharing of information and the public justification of why a probability was given, and tracking encouraged participants to outperform their previous track record and helped develop stronger teams of peers who could learn from one another.[83]

However, it is worth nothing that a large amount of time and resources go into selecting super-forecasting teams. The Good Judgement Project spent four years assembling their elite superforecasting team. It is difficult to imagine that such teams could be rolled out more extensively. Nevertheless, the approach itself is quite simple, and several people in the Existential Risk Studies community have attempted to adapt elements of it into their work. For example, just one hour of training in probabilistic reasoning noticeably improved forecasting accuracy.[84]

Whilst no superforecasters have attempted to predict the possibility of Human Extinction; the organisation Open Philanthropy commissioned a team of super-forecasters to predict the probability of a nuclear war.[85] Lessons from this approach could be incorporated by the Existential Risk Studies community in one of two ways. Firstly, it might be possible to train those who make Existential Risk predictions with the superforecasters' techniques. Secondly, applying more resources could motivate existing super-forecasting teams to make more relevant predictions of existential risk. Both will take considerable work, and it remains unclear how successful they will be.

Efforts to improve the quality of probability estimates at the point of solicitation focus on what questions are asked and how the person soliciting expert opinions engages with them. In a recent, albeit unpublished, solicitation of expert judgements of the probability of a Global Catastrophic Biological Risk, David Manheim used a variety of such approaches to solicit better quality information from experts in infectious disease.[86] He found that these experts were both poorly versed in probabilistic thinking and liable to reject the notion of a global catastrophic biological risk (in this case "a natural infectious disease that kills 1 billion people"); they disputed whether this could ever happen. However, Manheim proceeded to explain to these experts that other natural events could have catastrophic effects and challenged them to provide a fundamental reason why such an event was impossible within the infectious disease domain. By then, engaging the experts in scenario-based thinking about what properties such a disease would need to have, Manheim was able to solicit useful information with a reasonable degree of consensus between the experts. Only one expert continued to claim such an outcome was utterly impossible, but they now justified this claim, stating that this was a result of their belief that public health responses would always be sufficient to prevent such a pandemic. The experts remained unwilling to be quoted because they perceived a significant reputation risks in even discussing these extreme events. Post-solicitation methods focus on deliberation between experts, creating opportunities for experts to offer updated predictions or sometimes requiring them to adjust their judgements to move towards a consensus opinion. The most famous result is the so-called "Delphi technique" developed by RAND in the 1950s. This can be applied to a variety of foresight and horizon-scanning activities and uses a panel of experts who are asked to respond to a series of questions across two or more rounds. After each round, a facilitator provides an anonymised summary of the results, along with the reasons each expert provided for their answers. Extreme outliers must substantiate their position. Experts can then revise their judgements given the broader knowledge achieved through considering the responses of others leading, hopefully, to experts converging on the "correct" judgement. Delphi studies have been conducted to provide quantitative assessments in many areas of risk analysis, but the technique has not yet been used to provide quantitative estimates of existential risk, although Wintle et al. apply it to identify key emerging risks related to biotechnology in the Existential Risk context. [87] Other forms of structured expert elicitation that are related to, though not identical with, the Delphi technique have been harnessed to assess existential and Global Catastrophic Risk. For instance, Pamlin and Armstrong[88] used a complex multi-layered process of literature review, deliberative workshops and individual subjective judgements to select and assess Global Catastrophic Risks, but without multiple rounds of estimation. Another approach that has recently been developed, in part by Existential Risk researchers, is the IDEA (Investigate, Discuss, Estimate and Aggregate) protocol.[89] This drops the focus on seeking consensus and allows participants to discuss differences of opinion and defend probability estimates directly rather than responding to anonymised statements of reasons. The final independent estimates are given as anonymous submissions and then aggregated.

Whilst the Delphi technique and its relations aim to remove personal bias from predictions, as with all survey methods there may still be bias in the selection of the experts that can potentially lead to self-fulfilling prophecies.[90] Individual biases may influence people's willingness to update their judgment in light of evidence from the group and thus disproportionately sway the overall groups' findings. Moreover, some participants may wish to tailor their contributions to ensure that there is a concordant result, rather than rocking the boat with a contribution that throws the group further away in their estimate. Furthermore, the focus on consensus may be at the expense of cultural and other embedded differences in individuals' perspectives on information.[91] Finally, when the aggregation of expert opinion involves additional deliberation between experts, this can lead them to shift away from consensus and towards the most extreme views under discussion; individuals begin to cluster their identities around opposing positions, such as those defined along political or disciplinary lines.[92]

These techniques are relatively difficult to implement, requiring technical familiarity and the resources to convene a sufficient number of experts to implement them, but these barriers are lowering with time especially as the Delphi technique has a long track record of use in a variety of scientific and policy contexts. The fact that this approach can harness knowledge and expertise from across disciplinary backgrounds, requires individuals to substantiate judgements and encourages individuals to revise their first estimates in light of new information lends it a considerable degree of rigour, at least relative to many other methods that we have looked at. Whilst potentially controversial, the results are easy to communicate and are given credibility by the structured process through which they are obtained.

Bamber and Aspinall noted that experts in their study were "exceedingly uncertain about the answer to [the] key question".[93] They argue that whilst structured expert elicitation can help to quantify uncertainties; it does not overcome them. Such high degrees of uncertainty are often seen as prohibitive for quantitative research, and this may be part of why the Delphi Technique is often reserved for qualitative studies. However, we believe that this feature of the technique should be viewed in a very different light within the field of existential risk, where confidence in predictions is often overstated.

We believe that enhanced solicitation techniques have a significant underused potential to contribute to the quantification of existential risk. They are more rigorous, useful and able to handle uncertainty than individual or aggregated subjective opinions, although they are also harder to implement. A particular attraction of these techniques is their ability to open up a broad range of knowledge and perspectives on risk and to guide experts in combining this into coherent judgements. We rate these methods as High/Medium for rigour and Medium for uncertainty and utility, although we a higher degree of uncertainty for all three categories. However, we rated these techniques as only Low for accessibility due both to the time and expertise required to implement them well.

5.3 Prediction markets

Prediction markets function by providing a platform on which people can make trades based on their different assessments of the probability of an outcome or event. The more accurate any person is, the higher their payout. The price at which people are willing to make these trades depends on their probability assessments and their level of certainty in these assessments. This incentivizes individuals to be as rigorous and accurate as possible and allows for aggregation to take place over a potentially unlimited number of participants. The prediction market Metaculus uses trades with in-platform credits allowing individuals to perform actions such as posing their own questions. It has set up a market to establish the probability of human extinction,[94] although the market clearing price, which will represent its "final" prediction, will not be available until it closes in 2030. However, as Metaculus notes, this market, unlike its others, will not be able to pay out and users are therefore asked to make trades "in good faith" only. One proposal to overcome this barrier is to build the markets around trade in a resource that would help individuals survive a global catastrophe, such as access to survival shelters.[95]

Lionel Page and Robert Clemen argue that prediction markets are relatively well-calibrated when the time to expiration is relatively short, but that prices for the future are significantly biased.[96] One might overcome this barrier by establishing markets for events that would be related to, but not necessarily cause, an Existential Risk. For instance, prediction markets could be used to determine the probability that a large asteroid will pass within lunar orbit, that at least one nuclear weapon will be detonated by a non-state actor or some other "near miss" event that would help us understand Existential Risk without implying that humanity would actually go extinct.[97]

As with all markets, individuals who have limited information may assume that the market is better informed than they are and therefore not bid away from the current market price. This can cause price biases, where it becomes entrenched and prevents markets fully taking account of changing conditions. Nevertheless, prediction markets have proven success at making predictions even under situations of extreme uncertainty, such as whether CERN will locate the Higgs boson.[98]

Prediction markets currently have a strong track record, and there is considerable interest in their use, both amongst experts and as a means of "democratising" decision making. However, there are significant barriers to their application for Existential Risk. If a suitable platform could be established where participants were shown to have a clear interest in the long term, and their returns were guaranteed against inflation and loss of investment potential, perhaps via a philanthropic backer, then they might have an important part to play in assessing the probability of existential near misses. We rate prediction markets Low for uncertainty and utility and Medium for rigour and accessibility.

6. Discussion and Recommendations

In this section, we discuss the relative value of each of the methods that we have described above and make some recommendations for how they should be applied, implemented and evaluated by the Existential Risk Studies community.

6.1 Comparing methodologies

There are many methods currently being used, or with potential to be used, to quantify Existential Risk. Each method comes with its advantages and disadvantages, which we summarise in the following table.

Table 1: Comparing methods for quantifying existential risks.

Methodology	**Rigour**	**Uncertainty**	**Accessibility**	**Utility**	**Used for**
Analytical Approaches	Very Low	Very Low	High	Very Low	X-risk
Extrapolation and Toy Modelling	Low	Low	High	Low	Volcanoes, Pandemics, Nuclear, Space, Particle Physics, Asteroids
Fault Trees	Medium	Low	Medium*	Medium*	Nuclear, AI
Bayesian Networks	High / Medium	High	Low	Medium*	None
Adapting Large-Scale Models	High	High	Very Low	High	Pandemics, Climate, Asteroids
Individual Subjective Opinion	Low / Medium	Very Low	High	Low	X-risk, Climate, Asteroids, Nuclear

Methodology	**Rigour**	**Uncertainty**	**Accessibility**	**Utility**	**Used for**
Aggregated Opinion Surveys	Medium	Low	Medium*	Low	AI, Nuclear, X-risk
Weighted Aggregation	Low / Medium	Low	Low	Low	None
Enhanced Solicitation	High / Medium	Medium*	Low	Medium*	Pandemics, Nuclear, X-risk
Prediction Markets	Medium	Low	Medium*	Low	X-risk

There appear to be no standout "winners" from this analysis and every technique is rated Low on at least one criterion. The top scorers from our analysis as a whole are Bayesian networks, adapting existing models and Enhanced Solicitation Techniques, all of which score Low or Very Low in terms of accessibility. Of the more accessible approaches Toy Modelling and Aggregated Opinion Surveys perform best.

Given this variety of methodological virtues, we conclude that method selection should be understood in context and that the suitability of a method to a researcher's needs and circumstances is more important than its overall performance. At present, methodology choice seems to be strongly related to the nature of the studied threat. Some methods may well lend themselves to specific threats, depending on whether they have already been modelled at the sub existential level or whether there is a past historical record on which to build one's analysis. However, we feel that most of these methods could be applied far more widely and that more appropriate determinants of their use are the resources available to a team, whether the research is being undertaken for scientific or policy purposes and how findings are intended to be used.

Tonn and Stiefel go further and argue for giving the "results of all methods to a panel of experts to reflect upon before they are asked for holistic assessments".[99] However, that strikes us as potentially problematic because it leads to the homogenisation of quite diverse methodological perspectives and the potential loss of insight and introduction of bias that this entails. Realistically, it also represents a further loss of accessibility and therefore may put researchers off from conducting empirical studies to begin with. Hence, we conclude that it would be better to encourage researchers to focus on the methods that are best suited to their particular context and let a thousand flowers bloom.

6.2 Structured and deliberative approaches

We believe that the use of structured approaches, and especially enhanced solicitation techniques has been especially underdeveloped within the field of existential risk research and that this deserves more attention. While processes like the Delphi technique and superforecasting are not unproblematic, they have developed a good reputation in many scientific circles for being well suited to both interdisciplinary research and making judgements under uncertainty, two of the greatest challenges facing existential risk quantification.

In particular, two areas strike us as prime candidates for employing such techniques. Firstly, given the lack of a transparent methodology for establishing probabilities in the Pamlin and Armstrong report,[100] the Delphi or IDEAs technique may be an appropriate tool should the Global Challenges Foundation seek to update this research. Secondly, given the prevalence of unstructured surveys in the analysis of Artificial Intelligence as an existential threat, we believe that a more structured approach to combining expert opinions in this area would be valuable in providing a more rigorous perspective on a controversial subject.

6.3 Improving methodologies

Beyond this, however, our study serves to highlight the significant diversity in approaches to the implementation of these methods. There are examples of both good and bad practice in the literature at present and, regrettably, it is not always the good practice that is driving out the bad in the marketplace of ideas. In particular, many of the methods we considered allow researchers to objectively set out their reasoning process for others to critique and potentially update in light of new evidence and most have techniques for assessing and reporting degrees of uncertainty in a judgement. However, in very many cases such opportunities were not taken or were merely paid lip service despite requiring little, if any, additional effort.

The main reasons for not taking advantage of such opportunities are reputational. If one expresses uncertainty, then others are likely to see your judgements as less credible, and if one clearly sets out one's reasoning process, then others may see it as mistaken. These are not good reasons for bad science, and even if there is some argument to be made for simplification in public-facing communication, clear statements of methodology and uncertainty should be produced for the Existential Risk Studies community.

A good example is set by the IPCC who make use of a clear uncertainty framework in their reports. This combines probability judgments and confidence judgements, with separate terms used to describe each. For instance, terms such as "likely" present a probability judgment, whilst terms like "confident" are used to present degrees of certainty. According to the IPCC, authors guidance notes: "A level of confidence provides a qualitative synthesis of an author team's judgment about the validity of a finding; it integrates the evaluation of evidence and agreement in one metric."[101] A potential strength of this approach is that it can be sensitive to particular limitations within a domain, such as the availability of evidence, the level of disagreement about how to interpret that evidence, the robustness of models and methods that are currently used to evaluate it and the overall level of consensus that has been achieved.

Other good examples tend to be set by studies that come out of the "hard" sciences, including those relating to pandemics and space weather, or those that are embedded in risk analysis, such as the work of Anthony Barrett and Seth Baum on nuclear war. However, in each of these domains, there remain examples of bad, or even discredited, science that are still repeated by Existential Risk researchers, both in public-facing work and academic papers.

Conclusion

Despite the challenges involved, the quantification of existential risk seems highly likely to continue as a prominent strand of research in this area, for risk communication, research prioritisation and policy formation. We believe, however, that it is time that researchers in this field became more aware of how they can, and should, go about this process. There are a wide variety of methods that have been tried thus far, and none of these is definitively best, each having both merits and challenges. More importantly though, any of these approaches can be implemented well or badly and the mere fact that a certain probability assessment has been produced does not mean it is worthy of reproduction or inclusion in further analysis.

This is basic science and common sense. However, it is arguable that within the nascent field of existential risk research people have been insufficiently discriminating in this regard. This is not only problematic in that it risks using worse results when better ones are available; it also holds back the development of the field by failing to stimulate scholars to improve the quality of assessments that they produce.

(In the references that follow, bracketed mentions of 'source x' refer to the accompanying literature review contained in the online appendice, which is available at https://www.sciencedirect.com/science/article/abs/pii/S0016328719303313#sec0115)

https://www.sciencedirect.com/science/article/abs/pii/S0016328719303313#sec0115

Notes and References

[1] Sagan, C. 'Nuclear war and climatic catastrophe: Some policy implications', *Foreign Affairs, 62*(2) (1983): 257–92. https://doi.org/10.2307/20041818

[2] Barrett, A. M. 'Value of Global Catastrophic Risk (GCR) information: Cost-effectiveness-based approach for GCR reduction‘, *Decision Analysis, 14*(3) (2017): 187–203. https://doi.org/10.1287/deca.2017.0350; and Baum, S. and A. Barrett. 'Towards an integrated assessment of Global Catastrophic Risk', in *Catastrophic and Existential Risk: Proceedings of the First Colloquium*. Garrick Institute for the Risk Sciences, University of California (2017), pp.41–62.

[3] Sagan (1983).

[4] Currie, A. 'Existential risk, creativity and well-adapted science‘, *Studies in History and Philosophy of Science Part A* (2018). https://doi.org/10.1016/j.shpsa.2018.09.008

[5] Tonn, B. and D. Stiefel. 'Evaluating methods for estimating existential risks', *Risk Analysis, 33*(10) (2013): 1772–87. https://doi.org/10.1111/risa.12039

6 Avin, S., C. B. Wintle, J. Weitzdörfer, S. S. Ó hÉigeartaigh, W. J. Sutherland and M. J. Rees. 'Classifying global catastrophic risks', *Futures, 102* (2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001

7 Gott III, J. R. 'Implications of the Copernican principle for our future prospects', *Nature, 363* (1993): 315–19 (source 3).

8 Simpson, F. 'Apocalypse now? Reviving the Doomsday argument', *arXiv preprint arXiv:1611.03072* (2016) (source 4).

9 Bostrom, N. 'Are we living in a computer simulation?', *The Philosophical Quarterly, 53*(211) (2003): 243–55. https://doi.org/10.1111/1467-9213.00309

10 Bostrom, N. 'Where are they?', *Technology Review, 111*(3) (2008).

11 We are aware of one other paper that seeks to quantify existential risk using the Fermi Paradox, but the probability estimate is given in terms of an arbitrary free variable for the number of civilisations that have reached our level of development in our neighbourhood. Since it does not also attempt to quantify this variable, a final estimate cannot yet be produced, so we have not included this in our literature review. See Miller, J. D. and D. Felton. 'The Fermi paradox, Bayes' rule, and existential risk management', *Futures, 86* (2017): 44–57.

12 Leslie, J. *The End of the World: The Science and Ethics of Human Extinction*. Routledge (2002).

13 Wells, W. 'Human survivability', *Apocalypse When*? Springer Praxis Books (2009). https://doi.org/10.1007/978-0-387-09837-1_5. (source 4)

14 Bostrom, N. 'Dinosaurs, dodos, humans?', *Review of Contemporary Philosophy, (8)* (2009): 85–89. (source 46)

15 Decker, R. W. 'How often does a Minoan eruption occur?' *Thera and the Aegean world III, 2* (1990): 444–52 (source 50); Harris, B. 'The potential impact of super-volcanic eruptions on the Earth's atmosphere', *Weather, 63*(8) (2008): 221–25. https://doi.org/10.1002/wea.263. (source 51); Aspinall, W. et al. *GFDRR, Volcano Risk Study: Volcano Hazard and Exposure in GFDRR Priority Countries and Risk Mitigation Measures*. Bristol University Cabot Institute and NGI Norway for the World Bank: NGI Report 20100806 (2011) (source 52).

16 Lundgren, C. 'What are the odds? Assessing the probability of a nuclear war', *The Nonproliferation Review, 20*(2) (2013): 361–74. https://doi.org/10.1080/10736700.2013.799828 (source 14); Turchin, A. V. *Structure of the Global Catastrophe. Risks of Human Extinction in the XXI Century*. lulu.com (2008) (source 17).

17 Love, J. J. 'Credible occurrence probabilities for extreme geophysical events: Earthquakes, volcanic eruptions, magnetic storms', *Geophysical Research Letters, 39*(10) (2012). https://doi.org/10.1029/2012GL051431 (source 59); Homeier, N. et al. *Solar Storm Risk to the North American Electric Grid*. Lloyd's (2013) (source 60). Melott, A. L., B. S. Lieberman, C. M. Laird, L. D. Martin, M. V. Medvedev, B. C. Thomas ... and C. H. Jackman. 'Did a gamma-ray burst initiate the late Ordovician mass extinction?', *International Journal of Astrobiology, 3*(1) (2004): 55–61. https://doi.org/10.1017/S1473550404001910 (source 61); Gehrels, N., C. M. Laird, C. H. Jackman, J. K. Cannizzo, B. J. Mattson and W. Chen. 'Ozone depletion from nearby supernovae', *The Astrophysical Journal, 585*(2) (2003): 1169 (source 62).

18 Dar, A., A. De Rújula and U. Heinz. 'Will relativistic heavy-ion colliders destroy our planet?', *Physics Letters B, 470*(1–4) (1999): 142–48. https://doi.org/10.1016/S0370-2693(99)01307-6 (source 63); Jaffe, R. L., W. Busza, F. Wilczek and J. Sandweiss.

'Review of speculative "disaster scenarios" at RHIC', *Reviews of Modern Physics, 72*(4) (2000): 1125 (source 64); Tegmark, M. and N. Bostrom. 'Is a doomsday catastrophe likely?', *Nature, 438*(7069) (2005): 754–54. https://doi.org/10.1038/438754a (source 65); Ellis, J., G. Giudice, M. Mangano, I. Tkachev and U. Wiedemann. 'Review of the safety of LHC collisions', *Journal of Physics G: Nuclear and Particle Physics, 35*(11) (2008): 115004. https://doi.org/10.1088/0954-3899/35/11/115004 (source 66).

19 Hempsell, C. M. 'The investigation of natural global catastrophes', *Journal of the British Interplanetary Society, 57*(*1/2*) (2004): 2–13 (source 1); Synder-Beattie, A., T. Ord and M. Bonsall. 'An upper bound for the background rate of human extinction', *Scientific Reports* (2019) (source 2).

20 Yampolskiy, R. V. 'Predicting future AI failures from historic examples', *Foresight, 21*(1) (2018): 138–52. https://doi.org/10.1108/FS-04-2018-0034

21 Hanson, R. 'Catastrophe, social collapse, and human extinction', in N. Bostrom and M. M. Ćirković (eds.). *Global Catastrophic Risks*. Oxford University Press (2008), pp. 363–77.

22 Millett, P. and A. Snyder-Beattie. 'Existential risk and cost-effective biosecurity', *Health Security, 15*(4) (2017): 373–83. https://doi.org/10.1089/hs.2017.0028 (source 27 and 28).

23 Riley, P. 'On the probability of occurrence of extreme space weather events', *Space Weather, 10*(2) (2012). https://doi.org/10.1029/2011SW000734

24 Bagus, G. *Pandemic Risk Modelling*. Chicago Actuarial Association (2008) (source 23).

25 Day, T., J. B. André and A. Park. 'The evolutionary emergence of pandemic influenza', *Proceedings of the Royal Society B: Biological Sciences, 273*(1604) (2006): 2945–53 (source 20).

26 Fan, V. Y., D. T. Jamison and L. H. Summers. 'Pandemic risk: how large are the expected losses?', *Bulletin of the World Health Organization, 96*(2) (2018): 129. https://doi.org/10.2471/BLT.17.199588 (source 22).

27 Millet and Snyder-Beattie (2017).

28 Klotz, L. C. and E. J. Sylvester. 'The consequences of a lab escape of a potential pandemic pathogen', *Frontiers in Public Health, 2* (2014): 116. https://doi.org/10.3389/fpubh.2014.00116 (source 24)

29 Lipsitch, M. and T. V. Inglesby. 'Moratorium on research intended to create novel potential pandemic pathogens', *mBio, 5*(6) (2014). https://doi.org/10.1128/mBio.02366-14 (source 25)

30 Fouchier, R. A. 'Studies on influenza virus transmission between ferrets: the public health risks revisited', *MBio, 6*(1) (2015): e02560–14. https://doi.org/10.1128/mBio.02560-14 (source 26)

31 Hellman, M. 'Risk analysis of nuclear deterrence', *The Bent of Tau Beta Pi, 99*(2) (2008): 14 (source 12).

32 Currie, A. *Rock, Bone, and Ruin: An Optimist's Guide to the Historical Sciences*. MIT Press (2018).

33 Decker (1990).

34 Inglesby, T. V. and D. A. Relman. 'How likely is it that biological agents will be used deliberately to cause widespread harm? Policymakers and scientists need to take seriously the possibility that potential pandemic pathogens will be misused', *EMBO Reports, 17*(2) (2016): 127–30. https://doi.org/10.15252/embr.201541674

35 Fouchier (2015).

36 Love (2012).

37 Hellman (2008).

38 Ćirković, M. M., A. Sandberg and N. Bostrom. 'Anthropic shadow: Observation selection effects and human extinction risks', *Risk Analysis: An International Journal, 30*(10) (2010): 1495–1506. https://doi.org/10.1111/j.1539-6924.2010.01460.x

39 Manheim, D. 'Questioning estimates for natural pandemic risk', *Health Security, 16*(6) (2018): 381–90. Tegmark and Bostrom (2005).

40 Woo, G. 'Counterfactual disaster risk analysis', *Var. J.*, (2) (2018): 279–91.

41 Woo (2018).

42 Lundgren (2013). Lewis, P. M., H. Williams, B. Pelopidas and S. Aghlani. *Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy*. Chatham House, The Royal Institute of International Affairs (2014); Baum, S., R. de Neufville and A. Barrett. 'A model for the probability of nuclear war', *Global Catastrophic Risk Institute Working Paper, 18*(1) (2018).

43 Manheim (2018).

44 Barrett, A. M., S. D. Baum and K. Hostetler. 'Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia', *Science & Global Security, 21*(2) (2013): 106–33. https://doi.org/10.1080/08929882.2013.798984 (source 13). Whereas Barrett et al. model one set of nuclear war scenarios between Russia and the US, Baum et al. extend this fault tree to model all important nuclear war scenarios between all relevant states, but they stop short of quantifying the overall probability, so we have not included this in our literature review. See Baum, S., R. de Neufville and A. Barrett. 'A model for the probability of nuclear war', *SSRN Electronic Journal* (2018). http://www.doi.org/10.2139/ssrn.3137081

45 Baum, S., A. Barrett and R. V. Yampolskiy. 'Modeling and interpreting expert disagreement about artificial superintelligence', *Informatica, 41*(7) (2017): 419–28 (source 40).

46 Tonn and Stiefel (2013).

47 Khakzad, N., F. Khan, F. and P. Amyotte. 'Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches', *Reliability Engineering and System Safety, 96*(8) (2011): 925–32.

48 Bobbio, A., L. Portinale, M. Minichino and E. Ciancamerla. 'Improving the analysis of dependable systems by mapping fault trees into Bayesian networks', *Reliability Engineering and System Safety, 71*(3) (2001): 249–60.

49 Li, L., J. Wang, H. Leung and C. Jiang. 'Assessment of catastrophic risk using Bayesian network constructed from domain knowledge and spatial data', *Risk Analysis: An International Journal, 30*(7) (2010): 1157–75. https://doi.org/10.1111/j.1539-6924.2010.01429.x

50 Madhav, N. 'Modelling a modern-day Spanish flu pandemic', *AIR Worldwide* (February 21, 2013) (source 21).

51 Xu, Y. and V. Ramanathan. 'Well below 2 C: Mitigation strategies for avoiding dangerous to catastrophic climate changes', *Proceedings of the National Academy of Sciences, 114*(39) (2017): 10315–23. https://doi.org/10.1073/pnas.1618481114 (source 35)

52 NASA's Centre for Near-Earth Object Studies (CNEOS), https://cneos.jpl.nasa.gov/

(source 43); Harris, A. 'What spaceguard did', *Nature, 453*(7199) (2008): 1178 (source 47); National Research Council. *Defending Planet Earth: Near-Earth Object Surveys and Hazard Mitigation Strategies*. The National Academies Press (2010) (source 48).

53 Wagner, G. and M. Weitzman. *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton University Press (2015): 53–56 (source 32).

54 Atkinson, H., C. Tickell and D. Williams. *Report of the Task Force on Potentially Hazardous Near Earth Objects* (2000) (source 44).

55 Dunlop, I. and D. Spratt. *Disaster Alley: Climate Change Conflict and Risk*. Breakthrough — National Centre for Climate Restoration (2017) (source 34).

56 King, D., D. Schrag, Z. Dadi, Q. Ye and A. Ghosh. *Climate Change: A Risk Assessment*. Centre for Policy Research, University of Cambridge (2015) (source 33).

57 Pindyck, R. S. 'Climate change policy: What do the models tell us?', *Journal of Economic Literature, 51*(3) (2013): 860–72. https://doi.org/10.1257/jel.51.3.860

58 Aspinall, W. 'A route to more tractable expert advice', *Nature, 463*(7279) (2010): 294. https://doi.org/10.1038/463294a; Morgan, M. G. 'Use (and abuse) of expert elicitation in support of decision making for public policy', *Proceedings of the National Academy of Sciences, 111*(20) (2014): 7176–84. https://doi.org/10.1073/pnas.1319946111. However, note that Morgan specifically cautions against the use of subjective opinion for highly ambiguous phenomena, for which there may be no reliable expertise available.

59 Rees, M. J. *Our Final Century*. Basic Books (2003) (source 8).

60 Stern, N. et al. *Stern Review: The Economics of Climate Change* (Vol. 30) (2006). HM Treasury (source 11).

61 Bostrom, N. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9* (2002) (source 7).

62 Halstead, J. 'Stratospheric aerosol injection research and existential risk', *Futures, 102* (2018): 63–77. https://doi.org/10.1016/j.futures.2018.03.004 (source 36)

63 Chapman, C. R. 'The hazard of near-Earth asteroid impacts on Earth', *Earth and Planetary Science Letters, 222*(1) (2004): 1–15. https://doi.org/10.1016/j.epsl.2004.03.004 (source 45)

64 Tonn and Stiefel (2013).

65 In their paper Tonn and Stiefel illustrate what such reasoning might look like by producing a possible estimate for existential risk; however, we have decided not to include this in our literature review as it seems clear that they intended it only as an illustration rather than a serious attempt at probability estimation.

66 These figures are based on an illustration produced by the authors.

67 Hubbard, D. W. *How to Measure Anything: Finding the Value of Intangibles in Business*. John Wiley & Sons (2014).

68 Yudkowksy (2018).

69 Condorcet, M. D. *Essay on the Application of Analysis to the Probability of Majority Decisions*. Imprimerie Royale (1785); Galton, F. 'Vox populi (the wisdom of crowds)', *Nature, 75*(7) (1907): 450–51. https://doi.org/10.1038/075450a0

70 Tetlock, P. E., B. A. Mellers and J. P. Scoblic. 'Bringing probability judgments into policy debates via forecasting tournaments', *Science, 355*(6324) (2017): 481–83. https://doi.org/10.1126/science.aal3147

71 Müller, V. C. and N. Bostrom. 'Future progress in artificial intelligence: A survey of expert opinion', *Fundamental Issues of Artificial Intelligence*. Springer (2016), pp. 555–72 (source 38); and Grace, K., J. Salvatier, A. Dafoe, B. Zhang and O. Evans. 'When will AI exceed human performance? Evidence from AI experts', *Journal of Artificial Intelligence Research, 62* (2018): 729–54. https://doi.org/10.1613/jair.1.11222 (source 39).

72 'Experts see rising risk of nuclear war: Survey', *Project for the Study of the 21st Century* (2015) (source 15) https://www.scribd.com/document/289407938/PS21-Great-Power-Conflict-Report (source 15).

73 Sandberg, A. and N. Bostrom. 'Global Catastrophic Risks survey', *Technical Report #2008-1*. Future of Humanity Institute, Oxford University (2008): 1–5 (sources 9, 19, 30, 41, and 56).

74 Grace et al. (2018).

75 Turchin, A. 'Assessing the future plausibility of catastrophically dangerous AI', *Futures, 107* (2019): 45–58. https://doi.org/10.1016/j.futures.2018.11.007

76 Sandberg and Bostrom (2008).

77 Cooke, R. 'Experts in uncertainty: Opinion and subjective probability in science', *Oxford University Press on Demand* (1991). https://doi.org/10.5860/choice.29-5666

78 Colson, A. R. and R. M. Cooke. 'Expert elicitation: Using the classical model to validate experts' judgments', *Review of Environmental Economics and Policy, 12*(1) (2018): 113–32. https://doi.org/10.1093/reep/rex022; but see also Clemen, R. T. 'Comment on Cooke's classical method', *Reliability Engineering and System Safety, 93*(5) (2008): 760–65. https://doi.org/10.1016/j.ress.2008.02.003

79 Burgman, M. A., M. McBride, R. Ashton, A. Speirs-Bridge, L. Flander, B. Wintle ... and C. Twardy. 'Expert status and performance', *PLoS One, 6*(7) (2011): e22998. https://doi.org/10.1371/journal.pone.0022998

80 Bamber, J. L. and W. P. Aspinall. 'An expert judgement assessment of future sea level rise from the ice sheets', *Nature Climate Change, 3*(4) (2013): 424. https://doi.org/10.1038/nclimate1778

81 Tetlock et al. (2017).

82 Tetlock, P. E. and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Random House (2016). https://doi.org/10.1108/fs-12-2016-0061

83 Mellers, B., L. Ungar, J. Baron, J. Ramos, B. Gurcay, K. Fincher ... and T. Murray. 'Psychological strategies for winning a geopolitical forecasting tournament', *Psychological Science, 25*(5) (2014): 1106–15. https://doi.org/10.1177/0956797614524255

84 Chang, W., E. Chen, B. Mellers and P. Tetlock. 'Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments', *Judgment and Decision Making, 11*(5) (2016): 509–26. https://doi.org/10.1017/s1930297500004599

85 Good Judgment Project (non-public data, referenced by Carl Shulman: http://effective-altruism.com/ea/1rk/current_estimates_for_likelihood_of_xrisk/) (source 16).

86 Manheim, D. *Eliciting Rvaluations of Existential Risk from Infectious Disease*. Unpublished (2018). https://www.openphilanthropy.org/focus/global-catastrophic-risks/biosecurity/david-manheim-research-existential-risk (source 29).

87 Wintle, B. C., C. R. Boehm, C. Rhodes, J. C. Molloy, P. Millett, L. Adam ... and R. Doubleday. 'Point of View: A transatlantic perspective on 20 emerging issues in biological engineering', *Elife, 6* (2017): e30247. https://doi.org/10.7554/eLife.30247

88 Wintle, B. C., C. R. Boehm, C. Rhodes, J. C. Molloy, P. Millett, L. Adam ... and R. Doubleday. 'Point of View: A transatlantic perspective on 20 emerging issues in biological engineering', *Elife, 6* (2017): e30247. https://doi.org/10.7554/eLife.30247

89 Wintle, B. C., C. R. Boehm, C. Rhodes, J. C. Molloy, P. Millett, L. Adam ... and R. Doubleday. 'Point of View: A transatlantic perspective on 20 emerging issues in biological engineering', *Elife, 6* (2017): e30247. https://doi.org/10.7554/eLife.30247

90 Devaney, L. and M. Henchion. 'Who is a Delphi "expert"? Reflections on a bioeconomy expert selection procedure from Ireland', *Futures, 99* (2018): 45–55. https://doi.org/10.1016/j.futures.2018.03.017

91 Ahlqvist, T. and M. Rhisiart. 'Emerging pathways for critical futures research: Changing contexts and impacts of social theory', *Futures, 71* (2015): 91–104. https://doi.org/10.1016/j.futures.2015.07.012

92 Sunstein, C. R. 'Deliberative trouble? Why groups go to extremes', *The Yale Law Journal, 110*(1) (2000): 71–119. https://doi.org/10.2307/797587

93 Bamber and Aspinall (2013).

94 Metaculus Online Prediction Market. https://www.metaculus.com/questions/578/human-extinction-by-2100/ (source 10)

95 Hanson (2008).

96 Page, L., and R. T. Clemen. 'Do prediction markets produce well-calibrated probability forecasts?', *The Economic Journal, 123*(568) (2012): 491–513. https://doi.org/10.1111/j.1468-0297.2012.02561.x

97 We are grateful to Toby Ord for these suggestions.

98 Pennock, D. M., S. Lawrence, C. L. Giles and F. A. Nielsen. 'The real power of artificial markets', Science, 291(5506) (2001): 987–88. https://doi.org/10.1126/science.291.5506.987

99 Tonn and Stiefel (2013).

100 Pamlin and Armstrong (2015).

101 Mastrandrea, M. D., K. J. Mach, G. K. Plattner, O. Edenhofer, T. F. Stocker, C. B. Field,... & P. R. Matschoss. 'The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups', *Climatic Change, 108*(4) (2011): 675. https://doi.org/10.1007/s10584-011-0178-6

7. Scanning Horizons in Research, Policy and Practice

Bonnie C. Wintle, Mahlo N. C. Kennicutt II and William J. Sutherland

Highlights:

- Horizon scanning involves crowdsourcing information and drawing on communities of practice to sort, verify and analyse that information in order to look for early indications of poorly recognised threats and opportunities.
- "Exploratory" horizon scanning identifies novel issues by searching for the first signals of change, while "issue-centred" scanning monitors issues that have already been identified by searching for additional signals to confirm their emergence.
- The chapter assesses a range of horizon-scanning approaches and implementations, both manual and semi-automated, in relation to scope selection, input gathering, data sorting, cataloguing and clustering, result analysis and prioritisation, output utilisation, and process evaluation.
- If using manual approaches, structured methods are essential to mitigate biases that human horizon scanners are prone to. Manual approaches can be further improved by making the search for issues systematic. Semi-automated tools and AI will increasingly enable searches uninfluenced by the biases of the manual searcher.

https://doi.org/10.11647/OBP.0360.07

- Horizon scanning is most effective as an aid to policy-making when it is incorporated by organisations and decision-makers into the policy design process. There are a range of frameworks that can help translate scanning outputs into policy, such as road mapping.

This chapter surveys the development of horizon scanning within the conservation research community and its spread to other fields, including Existential Risk Studies. Since 2017, CSER has produced a range of horizon scans using some of the techniques described here and our researchers continue to work on developing them. Examples of the research undertaken within the field are discussed in more detail in Chapters 15 and 16 of this volume.

1. Introduction

Conservationists have long had to deal with a number of prominent, recurring issues, such as habitat loss and fragmentation, pollution, invasive species and wildlife harvesting, to name a few. On top of these well-known challenges, others have emerged. Over the last half century, these have included the impact of halogenated pesticides and defoliants, acid rain from coal-fired electricity generation, ecological impacts of biofuel production and atmospheric releases of ozone-depleting chemicals. In more recent times, concerns have emerged around microplastics and exploitation of the Arctic, although some changes also bring opportunities for conservation, such as using mobile phones to collect data. New and emerging issues tend to make policy and practice more difficult. They add to an already challenging agenda, and often require a response when knowledge of the problem is limited.

Emerging from the relatively new field of "futures" studies, horizon scanning is still developing as a method. By crowdsourcing information and drawing on communities of practice to sort, verify and analyse that information, horizon scanning offers an efficient way to look for early indications of poorly recognised threats and opportunities.[1] It aims to minimise surprises by foreseeing these threats and opportunities, enabling policy-makers and researchers to respond quickly to developing problems. Horizon scanning is an approach primarily used to retrieve, sort and organise information from different sectors that is relevant to the question at hand, in a similar process to intelligence gathering. It can also include varying degrees of analysis, interpretation and prioritisation, but deciding which issues to act on, and how to act on them, typically takes place after the horizon scanning, and is assisted by other "futures" tools, such as visioning, causal layered analysis, scenario planning and backcasting.[2] Recent frameworks have also been developed to link different futures tools, such as horizon scanning and scenario planning, together.[3]

Horizon-scanning outputs come in a wide range of forms. Some broadly describe a single trend that cuts across different parts of society, such as the rise of big data, or the future of a general area of interest, such as "Environmental Sustainability and Competitiveness".[4] These outputs are usually aligned with more general foresight programmes. Other exercises look at a set of more specific potential threats, such as invasive species that may arrive in the UK and threaten biodiversity,[5] and compare them in an approach similar to risk assessment. For the last 10 years, conservation scientists have run annual horizon scans to identify emerging issues with the potential to impact global conservation.[6] A similar approach has also been used to identify important scientific questions that, if answered, would help guide conservation practice and policy.[7]

As with any policy advisory work, there is always a risk that useful information is gathered but not followed up, as decisions are often driven by other, usually non-scientific, factors. This risk may be higher with unsolicited scans (grassroots scans produced by a community of practitioners, researchers or academics) rather than solicited scans (called for by policy- and decision-makers). It can be unclear where the responsibility lies for integrating outputs into policy-making, and uptake depends on the organisational culture at the time.[8] Schultz pointed to a conceptual contradiction between evidence-based policy and horizon scanning, where the latter searches for issues that may not be fully supported by a definitive body of evidence.[9] A more optimistic perspective is that horizon scanning needs to be embedded in a broader strategic foresight framework, to increase the likelihood that findings are translated into practice.[10] As mentioned above, horizon scanning *identifies* emerging and novel threats and opportunities as a first step, but other foresight tools serve different purposes along the pathway to adopting appropriate policy. These other foresight tools are not explicitly covered in this chapter, but we provide an example, *The Antarctic Science Scan and Roadmap Challenges Exercise*, of a hybrid horizon scanning activity where an accompanying road map was also produced to outline actionable recommendations (Box 2).

In this chapter, we introduce both general and specific approaches to horizon scanning, outline some ways of achieving and measuring impact, and explore how horizon scanning may evolve in the future.

2. Approaches to Horizon Scanning

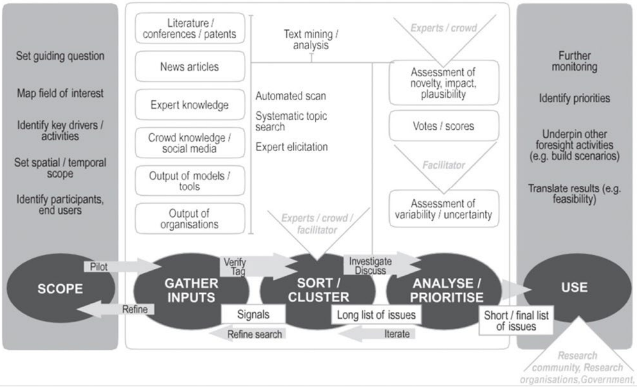

Set guiding question
Map field of interest
Identify key drivers / activities
Set spatial / temporal scope
Identify participants, end users
Literature / conferences / patents
News articles
Expert knowledge
Crowd knowledge / social media
Output of models / tools
Output of organisations
Text mining / analysis
Automated scan
Systematic topic search
Expert elicitation
Experts / crowd
Assessment of novelty, impact, plausibility
Votes / scores
Facilitator
Assessment of variability / uncertainty
Experts / crowd / facilitator
Further monitoring
Identify priorities
Underpin other foresight activities (e.g. build scenarios)
Translate results (e.g. feasibility)
Pilot
Verify
Tag
Investigate
Discuss
SCOPE
GATHER INPUTS
SORT / CLUSTER
ANALYSE / PRIORITISE
USE
Refine
Signals
Long list of issues
Short / final list of issues
Refine search
Iterate
Research community, Research organisations,Government,

Figure 1: General framework for horizon scanning, reflecting the key steps in the procedure (ovals), inputs and products (rounded rectangles), key outputs (rectangles), actors and end users (triangles), and activities and methods (floating text). Process adapted from Amanatidou et al.[11]

"Exploratory horizon scanning" identifies novel issues by searching for the first "signals" of change across a wide range of sources (such as an early scientific paper describing a potentially impactful new technology). "Issue-centred scanning" monitors issues that have already been identified by searching for additional signals that confirm or deny that the issue is truly emerging.[12] Signals can be organised into clusters (multiple pieces of information) that can either contribute to the evidence base around pre-identified issues, or form a long list of novel issues that are potentially emerging (Figure 1). The long list of issues can be further analysed and prioritised into a shortlist using methods detailed below. Some horizon scanning exercises take further steps to make the output more useful for the end user—for example, by assessing the policy relevance of the issues or the feasibility of addressing them, and by identifying those that warrant ongoing monitoring.[13]

There is a range of different ways to carry out horizon scanning; we introduce the main stages and provide some specific examples in the boxed texts and Table 1. Because our definition of horizon scanning concentrates largely on information retrieval, sorting and, to some extent, analysis and prioritisation, we focus here on methods that facilitate these activities.

Table 1: Approaches to horizon scanning (some activities and examples overlap)

Approach	**Examples**
Manual search of an invited expert group with Delphi-style prioritisation	Global conservation,[14] Antarctic science,[15] bioengineering,[16] Mediterranean conservation[17]
Manual search of a large crowd-sourced group (open call) with Delphi-style prioritisation (invited)	Future of the Illegal Wildlife Trade[18]
Automated open-source search and manual analysis/ prioritisation (usually by a community of experts)	IBIS,[19] Global Disease Detection Program (Centers for Disease Control and Prevention, www.cdc.gov/globalhealth/healthprotection/gdd/index.html), HealthMap (www.healthmap.org/en/), ProMed (www.promedmail.org/)
Advanced text analytics to identify emerging issues and research areas (e.g. sentiment analysis, machine learning)	FUSE Program (www.iarpa.gov/index.php/research-programs/fuse), Meta (https://meta.org/), X risk database (terra.cser.ac.uk)
Manual searches within an organisation (by employees, interns or volunteers), results tagged and catalogued in a database	US Forest Service,[20] UK Department for Environment, Food and Rural Affairs[21]

Approach	**Examples**
Comprehensive programme including scanning, sentiment analysis, scenario planning; manual and automated)	Singapore's Centre for Strategic Futures (www.csf.gov.sg/), partnered with the Risk Assessment and Horizon Scanning Programme Office
Expert opinion (voting, survey)	Global Risks Report 2019[22]
Regular meeting of a cross-disciplinary horizon-scanning group to discuss emerging issues and build database	Australasian Joint Agencies Scanning Network (www.ajasn.com.au/), Human Animal Infections and Risk Surveillance group (www.gov.uk/government/collections/human-animal-infections-and-risk-surveillance-group-hairs#risk-assessments-and-process)

2.1. Scoping

Like any major project, horizon scans need to be scoped and clear guidelines developed to assist scanners. A comprehensive scoping exercise addresses the following questions.

- What is the guiding question that defines what you want to know?
- How broadly or narrowly defined is the field of interest?
- What are the key drivers of change and activities in the field? It is common to organise thinking around a STEEP (Social, Technological, Economic, Environmental and Political factors) framework.
- What is the spatial scope? For instance, are you seeking issues with global or more localised impact?
- How far into the future should scanners be projecting?
- Who should be involved?
- Who are the potential end users?

Many of these considerations will be constrained by the resources available and the needs of the end user, but tools such as stakeholder analysis,[23] domain mapping[24] and issues trees[25] can be useful. Scoping exercises may also involve some pilot scanning to get a feel for how well-defined the task is. For example, preliminary scanning in a US Forest Service project that aimed to identify emerging issues that could affect forests and forestry in the future revealed that "natural resources and the environment" was too broad a topic for their exercise. Instead, it was narrowed to "forests".[26]

Horizon scans that rely heavily on people rather than computers to do the scanning reflect the biases of those participants. A well-structured procedure for obtaining judgements from participants (e.g. Figure 2) will go a long way to mitigating psychological biases,[27] but in order to capture a broad array of perspectives, involving a diverse group of people to identify and prioritise candidate issues is critical. A cognitively diverse group—comprising individuals who think differently—is thought to maximise collective wisdom and objectivity.[28] A good proxy for cognitive diversity is demographic diversity. Achieving demographic diversity can be challenging in practice. For example, there may be language barriers to overcome, and people with certain occupations (e.g. scholars) may be over-represented in horizon scans conducted by researchers. Inviting contributions from further afield, both geographically and from outside immediate peer circles, broadens the scope of issues considered. This might be achieved by putting out an open call for issues online and advertising it through relevant websites and email lists,[29] or posting a call for ideas on social media.

2.2. Gathering inputs

Inputs to a scan can either be gathered manually (by people) or with the aid of automated software, which is then (usually) analysed by people. Manual scanning typically involves a group of people monitoring current research and relevant trends (e.g. technology trends, disease trends or population trends) via desktop searches, attending conferences and consulting other people in their networks. Information can be manually scanned in news articles, social media, publications, grey literature and other output of relevant organisations (such as models and projections). This is typically the first step in a "Delphi-style" method that then goes on to analyse and prioritise candidate issues in a structured approach, usually involving one or more expert workshops (see Boxes 1 and 2 for examples and further descriptions of the procedure). Scanners could be provided with guidelines by a facilitator to direct their search, including suggestions of where to look. Manual methods have the advantage of accessing content that may not exist online (e.g. grey literature or unpublished research), or content that may be difficult to locate in the absence of known keywords to direct database and online searches. The downside of manual methods is that they are labour-intensive and may be exposed to the biases of the searcher, as they are less systematic.

Box 1: A Delphi-style method for horizon scanning in conservation.

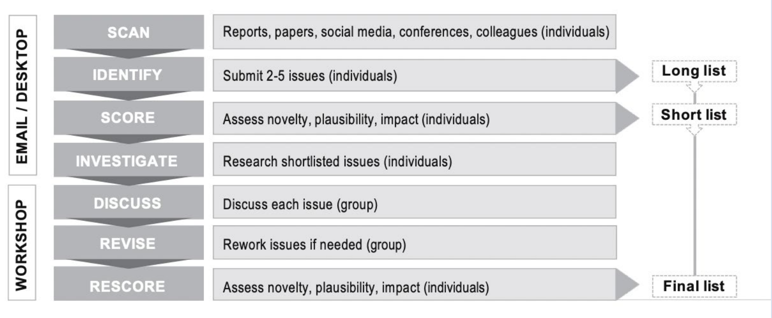

EMAIL / DESKTOP
WORKSHOP
SCAN
Reports, papers, social media, conferences, colleagues (individuals)
IDENTIFY
Submit 2-5 issues (individuals)
Long list
SCORE
Assess novelty, plausibility, impact (individuals)
Short list
INVESTIGATE
Research shortlisted issues (individuals)
DISCUSS
Discuss each issue (group)
REVISE
Rework issues if needed (group)
RESCORE
Assess novelty, plausibility, impact (individuals)
Final list

Figure 2: The Delphi-style horizon-scanning approach often used in conservation.[30] Figure reproduced from Wintle et al.,[31] published under the Creative Commons Attribution 4.0 Licence.

With its foundations in the Delphi Method,[32] this structured approach (Figure 3.2) was first applied in horizon scanning for conservation by Sutherland et al.[33] There are now several variants. The key features that make this approach "Delphi-style" are iteration (issues are submitted, scored, discussed and scored again) and anonymity of submissions and scoring. Typically, about 25 conservation experts from around the world participate in the following procedure. Over the course of several months, participants independently scan material from a variety of sources (e.g. papers, reports, websites, conferences) looking for issues (threats or opportunities) that are relatively novel, but that we should start planning for. Over email, each participant anonymously submits short summaries of two to five issues they have selected as the best "horizon-scanning" candidates, defined as reflecting a combination of novelty, plausibility and potential future impact on global conservation. The facilitator compiles the issue summaries and circulates them back to the group, who anonymously score each issue in terms of its suitability as a "horizon-scanning" item (using the definition above). A shortlist of the top-scoring issues, containing perhaps twice the total number sought, is recirculated back to participants. Each participant is assigned approximately five issues (not their own) to investigate further, gathering further evidence to support or oppose the issues' suitability. This means each issue will be cross-examined by at least two to three people. These five issues are usually assigned to people who are *not* considered experts in that subject matter, in the hope that they will have fewer preconceptions about the issue and that the experts will add their knowledge anyway. The whole group then meets at a workshop and systematically discusses each of the shortlisted issues (e.g. to consider new perspectives, relevant research, and whether the issue is genuinely novel or just a repackaging of an old issue). The issues are kept anonymous to reduce biases and allow for an open discussion. After the discussion, participants individually score the issues a second time. The top-scoring 15 are redrafted by one of the other group members and published each year in *Trends in Ecology & Evolution*.[34]

Box 2: Antarctic science scan and Roadmap Challenges project.

The international Antarctic community came together to horizon scan the highest priority scientific questions that researchers should aspire to answer in the next two decades and beyond. The approach included online submission of questions from the science research community, followed by a subset of 75 representatives (by nomination and voting) attending a workshop. At the workshop, approximately 1000 submitted questions were winnowed down to the 80 most important through methodical debate, discussion, revision and elimination by voting. All information used, including the 1000 submitted questions, was made publicly available in a database at a horizon scan website.[35] The horizon scan was followed by the Antarctic Roadmap Challenges project that was designed to delineate the critical requirements for delivering the highest priority research identified. The project addressed the challenges of enabling technologies, facilitating access to the region, providing logistics and infrastructure and capitalising on international cooperation. The process uniquely brought together scientists, research funders and those that provide the logistics for field research in the Antarctic. Online surveys of the community were conducted to identify the highest priority technological needs, and to assess the feasibility (time to development) and cost of these requirements. Sixty experts were assembled at a workshop to consider a series of topic-specific summary papers submitted by a range of Antarctic communities, survey results and summaries from the horizon scan, as well as existing documents addressing future Antarctic science directions, technologies and logistics requirements.[36]

Computer-assisted scanning is increasingly used for automating the process of gathering a vast quantity of inputs, often crowdsourced and usually from the internet.[37] Several such tools are now used in agriculture and health biosecurity to provide early detection of disease outbreaks (see Table 1 and Box 3 for examples).[38] Early online information, such as a tweet about a Tasmanian devil with a tumour on its face, or a YouTube video about a new device for targeting an invasive species, although unverified to begin with, may be critical for establishing the first in a series of signals that suggests a new or emerging threat.[39] Information on the internet can be retrieved in a number of ways. Keywords can be inserted into whole web search engines and/or particular websites can be targeted in more depth (e.g. Twitter can be searched using search terms, handles and hashtags). Research, news and current affairs can also be accessed via the RSS feeds of particular news and science sites, or by email and subscription to social media and blogs. Online data are often retrieved with the help of web scraping (accessing and storing particular web pages) and web crawling (accessing and storing links, and links of links from that page).[40] With the recent increase in "fake news", web searches require some form of quality control and vetting of sources: a process that can also be useful for *exposing* fake news. Large volumes of text scraped from the web, articles, patents, reports and other publications can be mined and filtered for potential relevance using automated software, such as machine learning algorithms.

Automated scanning is fast, systematic and comprehensive in its scope, but often relies on people—sometimes experts—to screen, review, and perhaps investigate all reports before on-posting or incorporating them.[41] For tools that scan across a wide range of topics, and those that use ongoing surveillance, this can be onerous and time-consuming. There are three other notable challenges to relying on online content for horizon scanning. First, material needs to already be posted on the web, and there may be a delay before an event, such as an invasive species incursion, is reported online. The second is that useful content is not always publicly available, as it can lie behind pay walls, be stored on intranets (e.g. grey literature), or secured because it is commercially, politically or personally sensitive. The third challenge is that most methods for obtaining online content rely on using the right keywords, which requires some idea of what you are looking for.

2.3. Sorting, cataloguing and clustering

Tagging and cataloguing content derived from both manual and automated scans (e.g. by relevance, credibility, source type, sectoral origin)[42] occurs concurrently with input gathering by scanners. Content can be further reorganised and vetted at a later stage. During this process, new search terms to direct further scanning can be generated, or existing search terms refined. Content can be organised according to a framework that also considers the level of response required and the strength of the evidence, which can help prioritise risks and other identified issues at a later stage.[43] Clustering methods, such as network analysis,[44] are useful for capturing cross-cutting issues that affect a number of topics of interest.

2.4. Analysing and prioritising

At this stage, a long list of issues will have been compiled, with some more suitable to the project aims than others. This can be an opportune time to reiterate objectives. Do you seek issues that most have not heard of? Do you intend to identify broad, developing topics or very specific developments (for example, the "increase in hydropower" versus "fragmentation effects of hydro-power in the Andean Amazon")? Are you interested in issues likely to arise soon or events that have a smaller probability of playing out in the long-term future? Does the output need to be useful to policy-makers? Many exercises, especially those with follow-up plans, aim to prioritise a select number of "most suitable" issues, and the precise manner in which such prioritisation decisions are made makes a real difference to the quality of the output.[45] Our experience with exercises that aim to identify novel issues is that participants gravitate towards well-known, although important, issues. Avoiding this requires strong chairing and a group that accepts the objective. To help overcome the problem, each participant can be asked whether they have heard of each issue, so that well-known topics can be excluded from the shortlist.

Box 3: Online horizon scanning: Intelligence-gathering for biosecurity

The International Biosecurity Intelligence System (IBIS) is a generic web-based application that focuses on animal, plant and marine health, and provides continuing surveillance of emerging pests and diseases, including environmental ones.[46] It also detects other environmental issues, such as harmful algal blooms. It is open source, in that it gathers articles from regular feeds of trusted sources (e.g. industry news, research) and publicly available online material, like news reports, blogs, published literature and Twitter feeds. Searches can be directed by broadly relevant keywords, such as “disease” or “outbreak” or “dead”, in addition to specific diseases of concern (e.g. “oyster herpes virus“). Articles can also be manually submitted by registered users to the application directly. A large expert community—the registered users, who are self-selected and approved by the administrator—then filter the articles, promoting those that they deem important and relevant to the home page, and demoting those that appear to be irrelevant or junk. Automated tools also assist with filtering (e.g. with machine learning and network cluster analysis), but as machine learning is still in its infancy, its use is limited to disease outbreaks from trusted sources. Items classified as junk by people are retained in a database to help the system’s artificial intelligence (AI) algorithms learn. The broader user community (anyone who signs up online) is alerted to items that have been flagged by the registered users as important, via a daily email new digest. IBIS is also “open analysis”, meaning that analysis of the publicly available information is performed openly by registered users. They can create or contribute to an emerging/ongoing issues dashboard that features a window for adding content, a Delphi-based forecasting section, links to related reports, share functions, comments and a map showing the location of events of interest (e.g. an outbreak). Registered users can also conduct their own searches and use integrated analytical tools to construct intelligence reports. IBIS has been effective for guiding policies and active risk management decisions for the Australian Government since 2006. The system may produce up to five Intel briefs a week on major issues affecting biosecurity and trade, allowing the government to respond to threats much faster than before. For instance, the system picked up a report of oyster herpes virus from a UK farm, which had previously purchased used aquaculture equipment from a disease-stricken oyster farm in France. Intelligence from IBIS revealed that businesses that had been closed down by the disease had been liquidating their equipment and selling to other countries. In response to this, the Australian Government changed its biosecurity policy to decontaminate all used aquaculture equipment on arrival.[47]

Within a manual Delphi-style approach (described in Boxes 1 and 2), issues are prioritised through an iterative scoring or voting process, usually facilitated online or in a workshop with a group of experts. The goal is to reduce a pool of potential horizon-scanning items or ideas to a smaller subset. The number of items, or issues, covered in the final list can vary, but tends to reflect around 10–30% of the initial items put forward.[48] As a point of comparison, the horizon scans described in Box 1 describe 15 issues annually, while the Antarctic hybrid horizon scan identified 80 shorter, priority scientific questions (Box 2). The final number may be constrained by how many the end user can realistically give their attention to (for a busy policy-maker, this may only be 15–20 half-page summaries), but is also driven by the number of (in)appropriate issues submitted. The main purpose of prioritisation is to remove issues that do not satisfy the selection criteria (novelty, plausibility, potential impact) and select those that are the most urgent or time sensitive. Prioritisation of issues will inevitably involve trade-offs, especially where different group members have different perspectives. Because individuals' diverging opinions can be masked in aggregated scores, analysing interrater concordance (e.g. with Kendall's *W*) affords insights into the level of agreement between contributors. In a diverse group, we would expect a wide variety of viewpoints to be voiced, but a core of shared opinions is often discernible.[49]

Items identified in a computerised scan (e.g. articles returned from a keyword search) are also prioritised by groups of people with varying levels of content expertise. People may be employed to sort through material, such as in governmental horizon-scanning programmes like in Singapore, or they may volunteer to do so because they are interested in the output, such as a farmer or epidemiologist concerned with news of disease outbreaks. Initially, items are sorted according to their relevance to the scanning aims (often done in the initial tagging/sorting process). Irrelevant items are discarded or moved to low priority. A second form of prioritisation involves flagging issues or topics that are particularly noteworthy.[50] This can be because signals have grown stronger (more evidence is gathered to suggest an issue is becoming a threat or presenting an opportunity for action),[51] or it might be because the potential consequences are so severe that the issue warrants immediate attention, even when evidence is limited or the probability is low ("wild cards").

2.5. Using the output

The previous step described prioritisation *within* the horizon scan to reduce a candidate set of issues. In that step, issues are ideally not judged according to importance, but rather according to less-subjective criteria, such as the likelihood of occurring or exceeding some threshold within a given timeframe. Prioritising which issues are the most important, and therefore should be acted on, is a different goal, and might be decided through follow-up, explicitly values-driven exercises involving representatives from government or relevant organisations.[52]

Bringing together a cross-section of policy-makers in a follow-up exercise can be useful, not only to identify those issues that require further monitoring or evidence before being acted on, but also to encourage prioritisation of cross-organisational issues, knowledge sharing, and collaborative development of policy. Ideally, feasibility assessments of the options available would be included (as carried out in the extension of the recent Antarctic scan, Box 2).

2.6. Evaluating the process

Assessing the success of horizon scans in identifying emerging issues is challenging, and has rarely been attempted. However, a recent review by Sutherland et al. examined the first of the annual global conservation scans described in Box 1[53] to consider how the issues identified in 2009 had developed.[54] This was assessed using several approaches: a mini-review was carried out for each topic; the trajectory of the number of articles in the scientific literature and news media that mentioned each topic in the years before and after their identification was examined; and a Delphi-style scoring process was used to assess each topic's change in importance. This showed that five of the 15 topics, including microplastic pollution, synthetic meat and environmental applications of mobile-sensing technology, appeared to have shown increased salience and effects. The development of six topics was considered moderate, three had not emerged and the effects of one topic were considered low.

As part of the same exercise, 12 global conservation organisations were questioned in 2010 about their awareness of—and current and anticipated involvement in—each of the topics identified in 2009.[55] This survey was repeated in 2018.[56] Awareness of all topics had increased, with the largest increases associated with micro-plastic pollution and synthetic meat; the change in organisational involvement was highest for microplastics and mobile-sensing technology. Perhaps the most surprising result was the number that had not heard of what are now mainstream issues: 77% for microplastics, 54% for synthetic meat and 31% for the use of mobile sensing technology. A decade ago the idea of collecting environmental data using phones was cutting-edge.

Thus, efforts have begun to examine the development of previously identified horizon-scan topics, but further research into the impact of horizon scans, and a consideration of issues that may have been "missed" (not identified but subsequently emerged as important) is needed.

3. Making a Difference With Horizon Scanning

Gauging the extent to which horizon-scanning outputs inform policy, future research directions and resource investments is not always straightforward and no-one has yet tested the effectiveness of this process. In instances where the primary decision-making organisation uses horizon scanning internally to assist with deliberations (e.g. scans to set priorities for a government agency), actions can be mapped directly against outcomes. In these cases, implementing the actions indicates impact. In other cases, scans can be driven by a community outside of government to set agreed future directions that can then be used to persuade external resource allocators. Even in cases where policy appears to reflect issues flagged in a horizon scan, it is difficult to trace direct influence, as inputs from multiple sources are often blended in final policy decisions without attribution. It also may take years for real-world impact to be realised. Nevertheless, there are ways in which uptake of horizon-scanning output can be encouraged.

As a starting point, horizon-scanning outputs can be matched to the organisations they are most relevant to. For example, policy-makers and practitioners can come together in a follow-up workshop to assess the importance of previously identified horizon-scanning issues for their organisation,[57] or the end user (e.g. policy-makers and practitioners) can be engaged in the horizon scan from the outset, as in a recent scan of research priorities for protected areas.[58] Similarly, horizon-scanning networks involving representatives from a range of government agencies, such as the Australasian Joint Agencies Scanning Network, or the UK Human Animal Infections and Risk Surveillance group, provide an ongoing forum for sharing information on new and emerging issues that potentially impact different departments and organisations. Regular meetings and reports are used to deliver this information to policy-makers in a timely way.[59]

In-depth follow-up analyses of horizon-scanning issues may also help policy-makers decide which to target first. A formal risk analysis of likelihood and consequences might be most appropriate for horizon-scanning outputs that compare similarly well-defined issues: for example, comparing one invasive species with another.[60] It may be more challenging if some of the issues in the candidate set are more coarse-grained than others (e.g. comparing ocean warming with a specific emerging fungal disease in some snakes). Nonetheless, risk-based prioritisation at least offers a framework for comparing and forecasting issues[61] and for formally considering the strength of evidence for each.[62]

Simply making horizon-scanning outputs known and available to policy-makers can encourage uptake. For example, issues identified in the annual global conservation scans (Box 1) have previously helped inform the UK's Natural Environment Research Council "Forward Look" strategic planning, but when a decision-maker does not already have a use in mind, it may be unclear what to do with horizon-scanning information without more context and guidance. Detecting signals and potential issues is only the first step towards making a difference: further intelligence about drivers is then needed to make sense of that information. For example, incorporating available data and modelling on air traffic movements with disease surveillance data might have helped anticipate the emergence of West Nile Virus in the United States in 1999.[63] It is the combination of horizon scanning, intelligence analysis (which provides context for the scanning output) and forecasting the chances of events unfolding that is particularly helpful in translating scanning outputs for policy-making. This can be embedded in a workflow, parts of which can be automated, such as compiling the context, narrative and structure into a digestible report on an important emerging issue (e.g. Box 3). When forecasting and open-analysis communities are already in place, this workflow can be delivered efficiently.[64]

Horizon scanning that occurs within organisations is evolving into a more effective tool than it was in its infancy. To facilitate the spread of best practice and reduce duplication, the UK has seen greater integration of horizon-scanning activities between different government departments, mainly in response to the Day Review.[65] The review recommended that horizon scans: (i) look beyond short-term agendas and parliamentary terms, (ii) focus on specific areas rather than broad topics in order to get more traction, (iii) are championed by those who use them in strategic decision-making, (iv) produce shorter outputs that are more likely to get the attention of senior decision-makers and (v) draw on inputs and existing analyses sourced from a "wide range of external institutions, academia, industry specialists and foreign governments". The extent to which all these recommendations have been implemented is unclear, but they represent a clear set of guidelines to follow.

There are a range of other useful frameworks that can be used for translating scanning outputs including road mapping the steps towards acting on different horizon-scanning issues, for example, by assessing the feasibility and estimating how long it would take to develop technologies needed to address particular research gaps (Box 2). The Antarctic science scan and roadmap has since been used to set National Antarctic Program goals, judge the effectiveness and relevance of past investments, and guide investment of other national programmes.[66]

4. Future Directions

We have discussed some of the pros and cons of different approaches to horizon scanning. If using a manual approach, structured methods are essential for mitigating the social and psychological biases that human horizon scanners are prone to, especially when forecasting complex and uncertain futures.[67] Although historically it has been criticised for confusing opinion with systematic prediction,[68] an iterative Delphi-style approach offers the advantage of drawing on the collective wisdom of a group, while affording individuals the opportunity to give private, anonymous judgements and revise them in light of information and reasoning provided by others. Compared with other elicitation approaches, such as traditional meetings, the Delphi method has also been found to improve forecasts and group judgements.[69] Manual approaches could be further improved by making the search for issues more systematic. Semi-automated tools and AI will increasingly enable searches uninfluenced by the biases of the manual searcher. For example, the Dutch "Metafore" horizon-scanning approach,[70] developed in The Hague Centre for Strategic Studies, already uses some automated approaches to systematically collect, parse, visualise and analyse a large "futures" database to complement their manual scanning.

Future horizon scanning and intelligence gathering may also see more open analysis, "citizen science" tools becoming adopted. While organisations are increasingly scanning open-source material (including news and social media), analyses typically remain internal.[71] This means the analyses are generally not available to external users in an unfiltered form or in a timely way, which is particularly important for risks such as disease spread. Governments may opt for confidentiality for both security and political reasons. For instance, negative public perceptions about a suspected emerging herpes virus in oysters might affect trade, which might delay the disclosure of this information by authorities, in turn delaying risk mitigation actions.[72] Intelligence tools (e.g. Box 3) that draw on a community of users to openly analyse news and information on potentially emerging issues offer more timely and transparent synthesis of information, which encourages more responsive decision-making. Examples of this can be seen in citizen science—for example, where citizen volunteers have helped analyse satellite-based information in the wake of natural disasters to help emergency responders to rapidly assess the damage.[73] In conservation science, involving a broader community of people in a participatory process like open analysis may also increase public support for science and the environment.[74] More open-source and open-analysis scanning tools in the future will also likely be complemented with better information visualisation and GIS (e.g. including maps that indicate where a relevant incident has taken place),[75] not only for identifying novel issues and monitoring issues that are already emerging, but also for locating and efficiently communicating this information.

Advanced text analytics, including text mining, will also provide a more comprehensive and systematic approach to future horizon scans. Indeed, some horizon-scanning centres, such as Singapore's Risk Assessment and Horizon Scanning programme, already use sentiment analysis—a way of computationally categorising subjective opinions expressed in text (e.g. positive, negative or neutral)—to uncover themes in content retrieved by their analysts. Even more sophisticated text analytics are becoming available, for example, to explore areas of disagreement, conflict or debate in the text of scientific literature to help track developments in science and technology.[76] They can also be used to detect language expressing excitement about a new idea, and other indicators of emergence, such as the increasing use of acronyms and abbreviations indicating that the scientific community is beginning to accept a technology or idea as established.[77] Through automation, new computational tools have the capacity to process a massive volume of papers and patents to anticipate which developments will have the biggest impact in the future.[78] These advances in text analytics have recently led to the development of a particularly powerful open-source AI tool, Meta (https://meta.org/), to help biomedical scientists and funders to connect emerging research areas and potential collaborators and inform investment. Due to the complexity of emerging issues (and complex environment for machines to learn in), progress towards detecting issues effectively through AI is slow. Computers may never outperform humans at natural language understanding, but steady improvements in the technology, coupled with the speed at which text can be processed by computers—in a range of languages—will undoubtedly add value to horizon scanning in the future.

Acknowledgements

Bonnie Wintle was supported by the Centre for the Study of Existential Risk at the University of Cambridge and funded by the Templeton World Charity Foundation (TWCF). William Sutherland is funded by Arcadia. Geoff Grossel provided useful insights into online scanning methods and applications.

Notes and References

1 van Rij, V. 'Joint horizon scanning: Identifying common strategic choices and questions for knowledge', *Science and Public Policy, 37* (2010): 7–18. https://doi.org/10.3152/030234210x484801; Sutherland, W. J. and H. J. Woodroof. 'The need for environmental horizon scanning', *Trends in Ecology & Evolution, 24* (2009): 523–27. https://doi.org/10.1016/j.tree.2009.04.008.

2 E.g. Glenn, J. C. and T. J. Gordon (eds.). *Futures Research Methodology—Version 3.0*. The Millennium Project (2009); Inayatullah, S. 'Futures studies: Theories and methods', in F. Gutierrez Junquera (ed.), *There's a Future: Visions for a Better World*. Banco Bilbao Vizcaya Argentaria Open Mind (2013), pp. 36–66. Cook, C. N., S. Inayatullah, M. A. Burgman et al. 'Strategic foresight: How planning for the unpredictable can improve environmental decision-making', *Trends in Ecology & Evolution, 29* (2014a): 531–41. https://doi.org/10.1016/j.tree.2014.07.005

3 Rowe, E., G. Wright and J. Derbyshire. 'Enhancing horizon scanning by utilizing pre-developed scenarios: Analysis of current practice and specification of a process improvement to aid the identification of important "weak signals"', *Technological Forecasting & Social Change, 125* (2017): 224–35. https://doi.org/10.1016/j.techfore.2017.08.001

4 Policy Horizons Canada. *Leading the Pack or Lagging Behind: A Foresight Study on Environmental Sustainability and Competitiveness*. Government of Canada (2011).

5 Roy, H. E., J. Peyton, D. C. Aldridge et al. 'Horizon scanning for invasive alien species with the potential to threaten biodiversity in Great Britain', *Global Change Biology, 20* (2014): 3859–71. https://doi.org/10.1111/gcb.12603

6 E.g. Sutherland, W. J., S. H. M. Butchart, B. Connor et al. 'A 2018 horizon scan of emerging issues for global conservation and biological diversity', *Trends in Ecology and Evolution, 33* (2018): 47–58. https://doi.org/10.1016/j.tree.2017.11.006

7 E.g. Sutherland, W. J., W. M. Adams, R. B. Aronson et al. 'One hundred questions of importance to the conservation of global biological diversity', *Conservation Biology, 23* (2009): 557–67. https://doi.org/10.1111/j.1523-1739.2009.01212.x

8 Delaney, K. and L. Osborne. 'Public sector horizon scanning-stocktake of the Australasian joint agencies scanning network', *Journal of Futures Studies, 17* (2013): 55–70.

9 Schultz, W. L. 'The cultural contradictions of managing change: using horizon scanning in an evidence-based policy context', *Foresight, 8* (2006): 3–12. https://doi.org/10.1108/14636680610681996

10 E.g. van Rij (2010); Cook et al. (2014a).

11 Amanatidou et al. (2012).

12 Amanatidou, E., M. Butter, V. Carabias et al 'On concepts and methods in horizon scanning: Lessons from initiating policy dialogues on emerging issues', *Science and Public Policy, 39*(2) (2012): 208–21. https://doi.org/10.1093/scipol/scs017

13 Sutherland, W. J., H. Allison, R. Aveling et al 'Enhancing the value of horizon scanning through collaborative review', *Oryx, 46*(3) (2012): 368–74. https://doi.org/10.1017/s0030605311001724

14 e.g. Sutherland et al. (2018).

15 E.g. Kennicutt, M. C., S. J. Chown, J. J. Cassano et al. 'A roadmap for Antarctic and Southern Ocean science for the next two decades and beyond', *Antarctic Science, 27*(1) (2015): 3–18. https://doi.org/10.1017/S0954102014000674

16 Wintle, B. C., C. R. Boehm, C. Rhodes et al. 'A transatlantic perspective on 20 emerging issues in biological engineering', *Elife, 6* (2017). https://doi.org/10.7554/elife.30247.e30247

17 Kark, S., W. J. Sutherland, U. Shanas et al. 'Priority questions and horizon scanning for conservation: A comparative study', *PLOS ONE*, *11* (2016). https://doi.org/10.1371/journal.pone.0145978

18 Esmail, N., B. C. Wintle, M. Sas-Rolfes et al. 'Emerging illegal wildlife trade issues in 2018: a global horizon scan', *SocArXiv* (25 April 2019). https://doi.org/10.31235/osf.io/b5azx

19 Grossel, G., A. Lyon and M. Nunn. 'Open- source intelligence gathering and open-analysis intelligence for biosecurity', in A. P. Robinson, M. Burgman, M. Nunn and T. Walshe (eds.). *Invasive Species: Risk Assessment and Management*. Cambridge University Press (2017), pp.84–92. https://doi.org/10.1017/9781139019606.005

20 Hines, A., D. N. Bengston, M. J. Dockry and A. Cowart. 'Setting up a horizon scanning system: A US Federal Agency example', *World Futures Review, 10*(2) (2018): 136–51. https://doi.org/10.1177/1946756717749613

21 Garnett, K., F. A. Lickorish, S. A. Rocks, G. Prpich, A. A. Rathe and S. J. T. Pollard. 'Integrating horizon scanning and strategic risk prioritisation using a weight of evidence framework to inform policy decisions', Science of The Total Environment, *560–61* (2016): 82–91. https://doi.org/10.1016/j.scitotenv.2016.04.040

22 World Economic Forum. *The Global Risks Report 2019: 14th Edition*. WEF (2019).

23 Reed, M. S., A. Graves, N. Dandy et al. 'Who's in and why? A typology of stakeholder analysis methods for natural resource management', *Journal of Environmental Management, 90* (2009): 1933–49. https://doi.org/10.1016/j.jenvman.2009.01.001

24 Lesley, M., J. Floyd and M. Oermann, M. 'Use of MindMapper software for research domain mapping', *Computers Informatics Nursing, 20* (2002): 229–35.

25 Government Office for Science. *The Futures Toolkit, Edition 1.0* (2017).

26 Hines et al. (2018).

27 Burgman, M. *Trusting Judgements: How to Get the Best Out of Experts*. Cambridge University Press (2015). https://doi.org/10.1017/CBO9781316282472

28 Page, S. E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press (2008).

29 E.g. Esmail et al. (2019).

30 Sutherland, W. J., E. Fleishman, M. B. Mascia et al. 'Methods for collaboratively identifying research priorities and emerging issues in science and policy', *Methods in Ecology and Evolution, 2* (2011): 238–47. https://doi.org/10.1111/j.2041-210x.2010.00083.x

31 Wintle et al. (2017).

32 Linstone, H. A. and M. Turoff. 'The Delphi Method: Techniques and applications', *Technometrics, 18*(3) (1976): 363. https://doi.org/10.2307/1268751;. Mukherjee, N., J. Huge, W. J. Sutherland et al. 'The Delphi technique in ecology and biological conservation: Applications and guidelines', *Methods in Ecology and Evolution* (2015). https://doi.org/10.1111/2041-210x.12387

33 Sutherland, W. J., M. J. Bailey, I. P. Bainbridge et al. 'Future novel threats and opportunities facing UK biodiversity identified by horizon scanning', *Journal of Applied Ecology, 45*(3) (2007): 821–33. https://doi.org/10.1111/j.1365-2664.2008.01474.x

34 e.g. Sutherland et al. (2018).

35 Kennicutt, M. C., S. J. Chown, J. J. Cassano et al. 'Polar research: Six priorities for Antarctic science', *Nature, 512* (2014): 23–25. https://doi.org/10.1038/512023a

36 Kennicutt et al. (2015).

37 Palomino, M. A., S. Bardsley, K. Bown et al. 'Web-based horizon scanning: Concepts and practice', *Foresight, 14* (2012): 355–73. https://doi.org/10.1108/14636681211269851

38 Salathé, M., L. Bengtsson, T. J. Bodnar et al. 'Digital epidemiology', *PLoS Computational Biology, 8*(7) (2012): e1002616. https://doi.org/10.1371/journal.pcbi.1002616;Kluberg, S., S. Mekaru, D. McIver et al. 'Global capacity for emerging infectious disease detection: 1996–2014', *Emerging Infectious Diseases, 22*(10) (2016). https://doi.org/10.3201/eid2210.151956 .

39 Grossel et al. (2017).

40 Hartley, D. M., N. P. Nelson, R. R. Arthur et al. 'An overview of internet biosurveillance', *Clinical Microbiology and Infection, 19* (2013): 1006–13. https://doi.org/10.1111/1469-0691.12273

41 Lyon, A. 'Review of online systems for biosecurity intelligence-gathering and analysis', *ACERA Project 1003*(6) (2010): 1097–1109.

42 e.g. Garnett et al. (2016); Hines et al. (2018).

43 Garnett et al. (2016).

44 Konnola, T., A. Salo, C. Cagnin et al. 'Facing the future: Scanning, synthesizing and sense-making in horizon scanning', *Science and Public Policy, 39* (2012): 222–31. https://doi.org/10.1093/scipol/scs021; Saritas, O. and I. Miles. 'Scan-4-Light: A searchlight function horizon scanning and trend monitoring project', *Foresight, 14* (2012): 489–510. https://doi.org/10.1108/14636681211284935

45 Sutherland, W. J. and M. Burgman. 'Policy advice: use experts wisely', *Nature, 526* (2015): 317–18. https://doi.org/10.1038/526317a

46 Grossel et al. (2017).

47 Burgman, M. 'Governance for effective policy-relevant scientific research: The shared governance model', *Asia & the Pacific Policy Studies, 2* (2015a): 441–51. https://doi.org/10.1002/app5.104

48 e.g. Kennicutt et al. (2014); Parker, M., A. Acland, H. J. Armstrong et al. 'Identifying the science and technology dimensions of emerging public policy issues through horizon scanning', *PLoS ONE, 9* (2014)., https://doi.org/10.1371/journal.pone.0096480; Kark et al. (2016); Wintle et al. (2017); Sutherland et al. (2018).

49 e.g. Wintle et al. (2017).

50 Grossel et al. (2017).

51 Cook, C. N., B. C. Wintle, S. C. Aldrich et al. 'Using strategic foresight to assess conservation opportunity', *Conservation Biology, 28* (2014b): 1474–83. https://doi.org/10.1111/cobi.12404

52 e.g. Sutherland et al. (2012).

53 Sutherland, W. J., M. Clout, I. M. Coˆte´ et al. 'A horizon scan of global conservation issues for 2010', *Trends in Ecology and Evolution, 25* (2010): 1–7. https://doi.org/10.1016/j.tree.2009.10.003

54 Sutherland, W. J., E. Fleishman, M. Clout et al. 'Ten years on: a review of the first global conservation horizon scan', *Trends in Ecology and Evolution, 34* (2019): 139. https://doi.org/10.1016/j.tree.2018.12.003

55 Sutherland et al. (2012).

56 Sutherland et al. (2019).

57 Sutherland et al. (2019).

58 Dudley, N., M. Hockings, S. Stolton et al. 'Priorities for protected area research', *Parks, 24* (2018): 35–50.

59 Delaney and Osborne (2013).

60 e.g. Roy et al. (2014).

61 Brookes, V. J., M. Hernandez-Jover, P. F. Black et al. 'Preparedness for emerging infectious diseases: Pathways from anticipation to action', *Epidemiology and Infection, 143* (2014): 2043–58.

62 Garnett et al. (2016).

63 Garmendia, A. E., H. J. Van Kruiningen and R. A. French. 'The West Nile virus: its recent emergence in North America', *Microbes and Infection, 3* (2001): 223–30. https://doi.org/10.1016/s1286-4579(01)01374-0; Brookes et al. (2014).

64 Grossel et al. (2017).

65 Day, J. *Review of Cross-Government Horizon Scanning*. Cabinet Office (2013).

66 National Academies of Sciences Engineering and Medicine. *A Strategic Vision for NSF Investment in Antarctic and Southern Ocean Research*. Author (2015). www.nsf.gov/funding/pgm_summ.jsp?pims_id=505320&org=OPP&from=home

67 Hanea, A. M., M. McBride, M. A. Burgman et al. 'Investigate discuss estimate aggregate for structured expert judgement', *International Journal of Forecasting, 33* (2017): 267–79. https://doi.org/10.1016/j.ijforecast.2016.02.008

68 Sackman, H. *Delphi Critique: Expert Opinion, Forecasting, and Group Process*. Lexington Books (1975).

69 Rowe, G. and G. Wright, G. 'Expert opinions in forecasting: Role of the Delphi technique', in J. S. Armstrong (ed.). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Kluwer Academic Publishers (2001), pp.125–44. https://doi.org/10.1007/978-0-306-47630-3_7

70 De Spiegeleire, S., F. van Duijne and E. Chivot. ’Towards Foresight 3.0: The HCSS Metafore Approach—A multilingual approach for exploring global foresights’, in T. Daim, D. Chiavetta, A. Porter and O. Saritas (eds.). *Anticipating Future Innovation Pathways Through Large Data Analysis. Innovation, Technology, and Knowledge Management*. Springer (2016). https://doi.org/10.1007/978-3-319-39056-7_6

71 Grossel et al. (2017).

72 Grossel et al. (2017).

73 Yore, R. ’Here’s how citizen scientists assisted with the disaster response in the Caribbean‘, *The Conversation* (18 October 2017).

74 Dickinson, J. L. and R. Bonney. *Citizen Science: Public Collaboration in Environmental Research*. Cornell University Press (2012). https://doi.org/10.7591/9780801463952

75 Dickinson, J. L., J. Shirk, D. Bonter et al. ‘The current state of citizen science as a tool for ecological research and public engagement’, *Frontiers in Ecology and the Environment, 10* (2012): 291–97. https://doi.org/10.1890/110236

76 Babko-Malaya, O., A. Meyers, J. Pustejovsky et al. ’Modeling debate within a scientific community‘, in *2013 International Conference on Social Intelligence and Technology*. IEEE (2013), pp.57–63). https://doi.org/10.1109/society.2013.18

77 Reardon, S. ‘Text-mining offers clues to success: US intelligence programme analyses language in patents and papers to identify next big technologies’, *Nature News, 509*(410) (2014).

78 Murdick, D. ’Foresight and Understanding from Scientific Exposition (FUSE): Predicting technical emergence from scientific and patent literature’, *IARPA*. US Office of the Director of National Intelligence (2015). www.iarpa.gov/images/files/programs/fuse/04-FUSE.pdf

8. Exploring Artificial Intelligence Futures

Shahar Avin

Highlights:

- This chapter provides a survey and initial categorisation of tools for exploring different futures for Artificial Intelligence, drawing mainly on work in the humanities.
- While no tools exist to reliably predict the future of AI, they can still help us expand our range of possible futures to reduce unexpected surprises and create common languages and models that enable constructive conversations about the kinds of futures we should try to occupy or avoid.
- Fictional narratives have long dominated thinking about AI futures but vary greatly in their degree of realism. They tend to suffer from a range of issues, including the need to entertain audiences, pressure to embody AI in physical forms like robots, a lack of diversity in authorship and representation, and limited accountability for their claims.
- Researchers from a variety of disciplines have produced high quality work drawing on insights from their fields. However, their predictions tend to fare poorly due to factors such as biases, partial perspectives, non-linear trends, and hidden feedback mechanisms. Furthermore, disagreement between experts can have a paralysing effect for audiences.

https://doi.org/10.11647/OBP.0360.08

- Group-based futures exploration can address some of these challenges using techniques like expert surveys, polling, interdisciplinary futures exercises, and expert elicitation. There are also opportunities to extrapolate futures from past and current data trends. Perhaps the most promising tools at our disposal are participatory futuring tools, including workshops, scenarios, and role-plays; however, these need to be realistic, integrative, and data-driven.

This chapter provides an alternative perspective on futures and foresight techniques, drawing more on work in the humanities, and their applications to Artificial Intelligence. The chapter's identification of high-quality scenario role-plays as an important methodological tool led directly to the development of Intelligence Rising (https://intelligencerising.org). The development of this tool is described in *Exploring AI Futures Through Role Play*[1] and is the subject of ongoing research, with further papers forthcoming. Group-based, collaborative and collective forms of knowledge generation and futures exploration are also discussed in Chapters 11 and 16.

1. Introduction

"Artificial Intelligence" (AI) is one of the more hyped-up terms in our current world, across academia, industry, policy and society.[2] The interest in AI, which long predates the current fascination, has given rise to numerous tools and methods to explore the potential futures of the technology, and its impact on human lives in a great variety of domains. While such visions are often drawn to utopian or dystopian extremes, more nuanced perspectives are also plentiful and varied, drawing on the history of the field, measurable progress and domain-specific expertise to extrapolate into possible future trends.

This chapter presents a survey of the different methods available for the exploration of AI futures, from narrative fiction in novels and movies, through disciplinary expert study of e.g. economic or philosophical aspects of AI futures, to integrative, interdisciplinary and participatory methods of exploring AI futures.

I begin in this section with setting common terms and boundaries for the discussion: the boundaries of "Artificial Intelligence" for the purposes of this chapter, certain contemporary technologies and trends that help ground and define the space of exploration, and an outline of the utopian and dystopian extremes that bound the current imagination of AI futures. I then go through each method of futures exploration in turn, providing a few examples and discussing some of the advantages and shortcomings of each. I conclude with a summary of the different methods and suggestions of strategies that may help furnish us with better information and expectations as we progress into a future shaped by AI.

1.1 Defining Artificial Intelligence

Given the newfound interest in AI, it is important to remember the history of AI as a field of research originating from work during the Second World War on computation and encryption, and the visions of the field's founders of machines that can learn and think like humans.[3]

While a precise definition of AI is elusive, I will satisfy myself with an analogy to artificial hearts and lungs: machines that can perform (some of) the functions of biological systems, in this case the human or animal brain/nervous system, while at the same time lacking other functions and often differing significantly in shape, material and other properties; this behavioural definition coheres well with the imitation game, or Turing test, that focuses on the machine "passing as" a human in the performance of a specific, delineated task within a specific, delineated domain. As the tasks become more vague, multifaceted and rich, and the domain becomes wider and less well defined, we move on the spectrum from narrow to general intelligence.[4]

The history of the field of AI research shows how wrong we tend to be, *a priori*, about which tasks are going to be easy, and which will be hard, for a machine to perform intelligently.[5] Breakthroughs in the field are often indexed to new exemplars of classes of tasks being successfully automated — for example, game playing[6] or image classification.[7]

1.2 Contemporary Artificial Intelligence

The current AI hype cycle is dominated by machine learning, and in particular by deep learning.[8] Relying on artificial neural networks, which emerged as broadly neurologically inspired algorithms in the second half of the 20th century,[9] these methods gained newfound success with the increasing availability of fast hardware and of large labelled datasets.[10]

In recent years we have seen increasing applications of deep learning in image classification, captioning, text comprehension, machine translation, and other domains. In essence, the statistically driven pattern recognition afforded by these technologies presented a sharp break from previous conceptions of AI as logic/rule-based, and a transition from the domain of explicit expert knowledge to domains of split-second recognition and response tasks (including, for example, driving-related tasks). However, the revolution also touched on expert domains that rely on pattern recognition, including medical image diagnosis[11] and Go game-play.[12]

Alongside these broadly positive developments, we have seen more ethically questionable applications, including in speech[13] and video synthesis[14] that mimics existing individuals, in learning to execute cyber attacks,[15] and in profiling and tracking individuals and crowds based on visual, behavioural and social patterns.[16] Existing and near future technologies enable a range of malicious use cases which require expanded or novel policy responses.[17]

1.3 Possible Artificial Intelligence futures

As we look further into the future, our imagination is guided by common tropes and narratives that predate the AI revolution.[18]

On the utopian end, super-intelligent thinking machines that have our interests as their guide, or with which we merge, could solve problems that have previously proven too hard to us mere humans, from challenges of environmental management and sustainability, to advanced energy sources and manufacturing techniques, to new forms of non-violent communication and new worlds of entertainment, to medical and biological advances that will make diseases a thing of the past, including the most terrifying disease of all — ageing and death.[19]

On the dystopian end, robotic armies, efficient and entirely lacking in compassion, coupled with the ability to tailor propaganda to every individual in every context on a massive scale, suggest a future captured by the power-hungry, ruthless few, with no hope of freedom or revolution.[20]

Worse still, if we ever create super-intelligent artificial systems, yet fail to align them with humanity's best interests, we may unleash a process of relentless optimisation, which will (gradually or rapidly) make our planet an uninhabitable environment for humans.[21]

The danger with extreme utopian and dystopian visions of technology futures is that they chart out what biologist Drew Endy called "the half pipe of doom",[22] a dynamic where all attention is focused on these extreme visions. More attention is warranted for mapping out the rich and complex space in between these extremes.

2. Exploring Artificial Intelligence Futures

We are not mere bystanders in this technological revolution. The futures we occupy will be futures of our own making, by action or inaction. To take meaningful action, we must come prepared with a range of alternatives, intervention points, a map of powerful actors and frameworks of critique. As the technical advances increasingly become widely accessible (at least on some level), it is our responsibility, as scholars, policy makers, and citizens, to engage with the technical literature and communities, to make sure our input is informed and realistic.

While it is the responsibility of the technical community to engage audiences affected by their creation (which, in the context of AI technologies, seems to be everyone), there is also a responsibility for those in the relevant positions to furnish decision-makers (again, broadly construed) with rich and diverse, yet fact-based and informed, futures narratives, maps and scenarios. Below I will survey a variety of tools available to us for exploring such futures, pointing out a few examples for each and considering advantages and limitations for each tool.

As a general note, this survey aims to be illustrative and comprehensive, but does not claim to be exhaustive. The examples chosen are by no means representative or exemplary — they are strongly biased by my regional, linguistic and disciplinary familiarity and preferences. Nonetheless, I hope the overall categorisation, and analysis of merits and limitations, will generalise across languages, regions and disciplines. I look forward to similar surveys from other perspectives and standpoints.

2.1 Fictional narratives

Probably the most widely recognised source of AI futures is fictional narratives, across different media such as print (novels, short stories, and graphic novels), music, films and television. These would often fall within the science-fiction genre, or one of its numerous sub-genres. A few examples, chosen somewhat carelessly from the vast trove of AI fictions, include Asimov's *Robot* series, Leckie's *Imperial Radch* trilogy, Banks' *Culture* novels, Wells' *Murderbot Diaries* series, *The Jetsons*, the *Terminator* franchise of movies and TV series, the movie *Metropolis*, and the musical concept series of the same name by Monáe.

Works vary greatly in their degree of realism, from those rich in heavily researched details, to those that deploy fantastical technology as a tool to explore some other topic of interest, such as emotions, power relations, agency or consciousness. As such, fictional AI narratives can be both a source of broadened horizons and challenging ethical questions, but also a source of harm when it comes to exploring *our* AI futures — they can anchor us to extreme, implausible or misleading narratives, and, when they gain widespread popularity, can prevent more nuanced or different narratives from gaining attention.

The challenge for fictional AI narratives to provide useful guidance is further aggravated by four sources: the need to entertain, the pressure to embody, a lack of diversity, and a limited accountability.

2.1.1 The need to entertain

Authors and scriptwriters need to eat and pay rent, and the amount of remuneration they receive is linked to the popularity of their creations, either directly through sales or indirectly through the likelihood of contracting. Especially with high-budget production costs, e.g. in Hollywood films,[23] scripts are likely to be more popular if they elicit a positive response from a broad audience, i.e. when they entertain. There is no *prima facie* reason to think that what makes for good entertainment also makes for a useful guide for the future, and many factors are likely to point to these two coming apart, such as the cognitive load of complexity and other cognitive biases,[24] or the appeal of extremes.[25]

2.1.2 The pressure to embody

Especially in visual media, but also in written form, narratives are made more accessible if the AI technologies discussed are somehow concretised or embodied, e.g. in the form of robots, androids, cyborgs or other machine bodies.[26] Such embodiment serves as a useful tool for exploring a range of pertinent issues, but also runs the risk of distracting us from other forms of intelligence that are less easy to make tangible, such as algorithms, computer networks, swarm intelligence and adaptive complex systems. The pressure to embody relates to, and is made complicated by, the proliferation of embodied instances and fictions of Artificial Intelligence, either as commercial products[27] or as artistic creations of robots and thinking machines in visual and physical forms — for example, robot toys or the illustrations that accompany news articles and publications. In general, as per my definition in the beginning, our understanding of Artificial Intelligence should focus on action and behaviour rather than form, though there are good arguments suggesting the two are linked.[28]

2.1.3 Lack of diversity

While narrative fictions may well provide us with the most rich and diverse exploration of possible AI futures, we should be mindful that not all identities and perspectives are represented in fictional narratives, and that the mere existence of a work does not readily translate into widespread adoption; narratives, like individuals, groups and world views, can be marginalised. While science fiction has been one of the outlets for heterodox and marginalised groups to make their voices heard,[29] this is not universally welcome,[30] and the distribution of attention is still heavily skewed towards the most popular works.[31]

2.1.4 Limited accountability

Creators of fictional narratives receive feedback from two main sources, their audience (through purchases and engagement with their works) and their critics. While these sources of feedback may occasionally comment or reflect on a work's ability to guide individuals and publics as they prepare for the future, this is not seen as a main aim of the works not an essential part of it.[32] In particular, there is little recognition of the possible harms that can follow misleading representations, though it is reasonable to argue that such harms are limited, especially in the absence of better guidance, and the fact that experts deliberately aiming to provide such guidance tend to fare quite poorly (Armstrong and Sotala, 2015).

2.2 Single-discipline futures explorations

As part of the phenomenon of AI hype, we are seeing an increase in the number of non-fiction books exploring the potential implications of Artificial Intelligence for the future, though of course such books have been published since before the field became established in academia, and previous "AI summers" have led to previous periods of increased publication. The authors who publish on the topic come from a wide range of disciplines, and deploy varying methods and arguments from diverse sources. These contribute to a richer understanding of what is, at heart, a multifaceted phenomenon.

For example, AI researchers spend just as much time on the history and sociology of the field, and on dispelling misconceptions, as they do on laying down observations and arguments with relevance for the future; [33] mathematicians and physicists focus on the world as seen through the lens of information, models and mathematics, and the AI futures that such a perspective underwrites; [34] technologists focus on underlying technology trends and quantitative predictions;[35] risk analysts explore the various pathways by which AI technologies could lead to future catastrophes;[36] economists focus on the impacts of AI technologies on the economy, productivity and jobs;[37] self-published, self-proclaimed business thought-leaders share their advice for the future;[38] political commentators write manifestos arguing for a particular future;[39] and philosophers examine the very nature of intelligence, and what happens when we extrapolate our understanding of it, and related concepts, into future capabilities that exceed what evolution has been able to generate.[40]

While the quality of research and arguments presented in such works tends to be high (as academic and public reputations are at stake), any predictions presented in such works tend to fare poorly, due to numerous factors including biases, partial perspectives, non-linear and discontinuous trends, hidden feedback mechanisms, and limited ability to calibrate predictions.[41] Furthermore, disagreement between experts, while to be expected given the uncertainties involved, can have a paralysing effect for audiences, a fact that can be exploited.[42]

If fictional narratives are best seen as a rich and fertile ground for futures imagination (as long as we do not get too distracted by the flashy and popular), expert explorations provide a rich toolset of arguments, trends and perspectives with which we can approach the future with an informed, critical stance, as long as we appreciate the deep uncertainty involved and avoid taking any trend or prediction at face value.

2.3 Group-based futures exploration

The nature of the problem being addressed — what are possible AI futures and which ones we should aim for or avoid (and how) — is inherently complex, multi-faceted and interdisciplinary. It is therefore natural to explore this problem through utilising diverse groups. There are various methods to do this, each with advantages and disadvantages (Rowe and Beard, 2018).

2.3.1 Expert surveys

What do different individuals think about the future of AI? One way to find out is to ask them. While survey design is not an easy task, we have the ability to improve upon past designs, and regularly update our questions, the target community, and the knowledge on which they draw (as more experience is gained over time).

Surveys amongst experts have been used in particular to explore questions of timing and broad assessment of impact — when will certain capabilities become available, and will they have a positive or negative impact?[43] As surveys only tell us *what* people think, rather than *why* they think it, they are best treated not as a calibrated prediction of the future (as all estimates could be flawed in the same way), but rather a useful data point about what beliefs are prevalent right now, which in itself is useful for exploring what beliefs might hold currency in the future, and how these might affect the future of AI.

2.3.2 Public polling

Public polling aims to examine both public understanding of the technology, the desirability of possible applications and concerns about possible uses and misuses of the technology.[44] While it may be tempting to interpret these polls as "hard data" on public preferences, it should be remembered that many factors affect responses.[45] In the Royal Society study cited above, conducted by Ipsos Mori, poll findings were compared with surveys of focus groups that had in-depth interactions with experts and structured discussions around the survey questions. Such practices bring polling closer to participatory futures workshops, discussed below.

2.3.3 Interdisciplinary futures studies

Often, we would want to go beyond an aggregate of single points-of-view, aiming for a more holistic understanding of some aspect of the future of AI through interactions between experts. Such interactions can be one-off or long standing, and they can be more or less structured (Rowe and Beard, 2018). An example of a broad-scoped, long-term academically led interdisciplinary study is the Stanford

100-year study of Artificial Intelligence.[46] An example of a more focused study is the workshop that led to the report on the potential for malicious use of Artificial Intelligence.[47] While such studies offer a depth advantage over surveys, and a diversity advantage over single-domain studies, they still face challenges of scope and inclusion: too narrow focus, on either topic or participants, can lead to a narrow or partial view, while too broad scoping and inclusion can make the process unmanageable.[48]

2.3.4 Evidence synthesis and expert elicitation

With a growing evidence base relevant to AI futures, policy-making and policy-guiding bodies are beginning to conduct structured evidence synthesis studies.[49] The methodologies for conducting such studies have been improved over the years in other evidence-reliant policy domains, and many lessons can be ported over, such as making evidence synthesis more inclusive, rigorous, transparent and accessible.[50]

We are also seeing efforts from governments to solicit expertise from a broad range of source, as early fact-finding steps that could lead to or inform policy in this space.[51] While such efforts are welcome — both in their interdisciplinary and participatory nature — through their democratic mandate, and through the proximity of expertise and accountable decision making, it should be noted that results still very much depend on the experts in the room, that such exercises tend to avoid areas of high uncertainty or disagreement (which may be the areas demanding most attention), and that the issues are often global and open in nature, limiting the effectiveness of national strategy and regulation.

2.4 Extrapolating from past and current data trends

While historical trends may provide only a limited guide to the future when it comes to emerging technologies,[52] it is still useful to have an up-to-date understanding of the state-of-the-art, especially when the field is progressing at a rapid pace leaving many outside the cutting edge with an out-dated view of what contemporary capabilities are (and are not). This is a constructive and interdisciplinary effort, as the tools to measure performance of AI technologies are just as much in flux as the technology itself. Measurements of the technology focus either on performance[53] or the resource use of the technology in terms of data or compute,[54] though other dimensions could also be measured.[55] Other efforts go beyond the technology itself and also track the ecosystem in which the technology is developed, looking at hardware, conference attendance numbers, publications, enrolment, etc.[56]

2.5 Interactive futures narratives and scenarios

For most of the futures exploration tools described above, the audience is passive, and is being communicated at via text or vision and sound. Even surveys of the public often involve only localised and limited contributions from each individual. However, there also exist tools that enable the audience to take a more active role, either in a pre-defined narrative or in the co-creation of narratives. The emphasis on greater public participation is a key tenant of responsible research and innovation[57] and it applies with force to the field of Artificial Intelligence.[58]

2.5.1 Participatory futures workshops

On the more formal end, participatory future workshops,[59] or one of the numerous variations on the theme,[60] go through a structured engagement between different stakeholders. These reflect the (originally more corporate and less open) processes of scenario planning.[61] Similar to scenario planning, where participants explore a range of possible futures as a team, wargaming[62] and drama theory[63] use role-play to place participants in opposing roles, to explore what strategies may emerge or investigate novel opportunities for cooperation and resolution. While the author knows of no such exercises on long-term AI futures, nearer-term exercises — for example, on autonomous driving — are already taking place.[64] When such exercises have the support of government and buy-in from both experts and non-experts, they can prove to be highly valuable tools in preparing for AI futures; indeed, they come close to certain visions of the ideal interaction between science and society.[65] However, they also require significant resources and expertise to carry out well.

2.5.2 Interactive fictions

At the less participatory end, but still allowing the audience to play a more active role, are interactive fictions, especially in the medium of video games. While Artificial Intelligence, as a long-standing science fiction trope, has been depicted in video games for decades, recent games incorporate more of the nuanced arguments presented about the potential futures and characteristics of AI.

For example, *The Red Strings Club* explores fundamental questions of machine ethics in an interactive dialogue with the player,[66] and *Universal Paperclips* allows the player to experience a thought experiment created to explore the "orthogonality thesis",[67] the argument that arbitrarily high levels of intelligence are compatible with a wide range of ultimate goals, including ones that would seem to us foolish or nonsensical.[68]

Other video games focus less on the narrative element, but rather present a rich simulator in which Artificial Intelligence is one of many technologies available to the player, allowing the exploration of a wide range of future AI scenarios and their interplay with other systems such as diplomacy or resource management. Examples include *Stellaris*,[69] in which Artificial Intelligence technologies are available to the player as they establish their galactic empire, or the *Superintelligence* mod[70] for *Sid Meier's Civilisation V*,[71] which allows the player, in the shoes of a world leader, to gain a strategic advantage using AI and achieve a scientific victory by creating an Artificial Superintelligence, while risking the creation of an unsafe superintelligence which could lead to an existential catastrophe.

2.5.3 Role-play scenarios

While video games allow audiences to take a more active role in the exploration of possible AI futures within the game environment, they hardly satisfy the call for public participation in jointly imagining and constructing the future of emerging technologies. To explore AI futures in a collaborative and inclusive manner, experts and audiences must explore them together. One way to achieve this is through the joint exploration of narratives in role-play games.

Scenarios that have been developed with expert participation through any of the methods above, or through other means, can be circulated more broadly as templates for role-play games amongst interested parties. At the hobbyist level, game systems such as *Revolt of the Machines*[72] and *Mutant: Mechatron*[73] allow players to collectively explore a possible AI future. While these are often very entertaining, they may fall into the same failures as narrative fictions. It seems there is currently an unmet need for realistic and engaging AI futures role-play game systems.

3. Summary and Conclusion

As AI hype drives utopian and dystopian visions, while rapid technological progress and adoption leaves many of us uncertain about the future impacts on our lives, the need for rich, informative, and grounded AI futures narratives is clear. It is also clear that there is a wide range of tools to develop such narratives, many of which are available to creators and experts outside the AI research community. It is less clear, however, how best to utilise each of the available tools, with what urgency and in which domains. The table below summarises the different tools surveyed above, with their respective advantages and limitations.

Table 1: Tools to develop Artificial Intelligence futures narratives.

Tool	Existing abundance	Skills and resources required	Advantages	Limitations
Fictional narratives	Overly abundant	Creative writing, production costs for film	Unbridled imagination, (relatively) open participation	Lack of realism, pull to extremes, lack of accountability, lack of diversity, skewed popularity distribution
Single-discipline futures exploration	Growing rapidly, though some disciplines are still missing	Domain expertise, familiarity with AI, forecasting skills	Deep dives into relevant facts and arguments	Predictive power is poor, disagreements can paralyse, not easy to integrate across disciplines
Surveys	Few key studies	Survey design, resources to carry out the survey	Aggregate evidence can counteract some biases, present a snapshot of current beliefs	Survey design is hard, topic in flux, misunderstanding is commonplace; poor predictive power
Interdisciplinary futures exploration	Few but growing rapidly	Interdisciplinary facilitation, network of stakeholders, time and geographic availability	Holistic view of complex topics, opportunity to directly engage with policy-makers and other key stakeholders	Risk of groupthink, conservatism; scoping is difficult: too narrow and miss opportunities and challenges, too broad and becomes intractable
Evidence synthesis	Few	Access to studies in a range of disciplines and expertise to assess them and communicate findings	Evidence-based holistic picture drawing on a wide range of works, prepared with policy in mind	Time and labour intensive, evidence may be partial and rapidly changing, best practices still evolving

Tool	**Existing abundance**	**Skills and resources required**	**Advantages**	**Limitations**
Extrapolating data trends	Few key hubs, abundant but disperse data	Familiarity with the field and the techniques of AI, measurement platforms, data harvesting and curation	Historical and contemporary measurements can be largely uncontested, informative	Difficult to extrapolate from past trends due to non-linearity, feedback, potential for discontinuity; need to constantly evolve and adapt measurements
Participatory futures workshops	None on long-term AI, few on short term issues such as self-driving cars	Buy-in from experts and non-expert participants, budget for workshops, facilitation skills, time of participants	Participatory, expert-informed exploration of future scenarios, legitimacy for policy guidance	Difficult to get buy-in and time commitment from experts and stakeholders, requires significant investment to tutor non-experts
Interactive fictions	Several, though few with realistic representations informed by recent advances	Game development skills and budget	Audience takes an active role, can explore alternatives, simulators offer a combinatorial explosion of options	Similar to fictional narratives, plus limitations of what can be represented effectively with limited skills and budget
Role-play scenarios	Few	Facilitation, game/ scenario design	Stakeholders can come together to co-explore possible futures	Information gaps in the group can slow down or derail the conversation, strongly depends on the available expertise and facilitation skills

As can be expected, no tool is strictly better than all other tools. Some provide more evidence-based, deep analysis, but tend to be limited in the range of questions they can cover and place barriers on participation. Others allow for more diverse and integrative perspectives, but tend to preclude detailed and in-depth analysis or come at a very high cost in terms of time and facilitation. Instead of judging individual futures narratives in isolation, it may be more useful to look at the entire ecosystem of future AI narratives, asking whether certain narratives are dominating our imagination without sufficient warrant, or if there are tools and narratives that are underutilised or gaining insufficient attention. At present, it seems that not enough attention is being given to data-driven, realistic, integrative, and participatory scenario role-plays, which can build on and integrate a range of other tools and narratives and make them more accessible to a wider audience in a more nuanced way. A more balanced portfolio is called for.

As we act to critique and curate the ecosystem of AI futures, we should keep in mind the aims of these narratives: beyond entertainment and education, there are real ongoing processes of technological development and deployment that currently have, are likely to continue to have, significant social impacts. These processes are not isolated from the societies in which they take place, and the interactions between technology developers, policymakers, diverse stakeholders and numerous publics are mediated and shaped by the futures narratives each group has access to. Thus, AI futures narratives play a crucial role in making sure we arrive at futures of our own choosing, that reflect our values and preferences, that minimise frictions along the path, and that do not take us by surprise. Thus, critique and curation of AI futures is an integral part of the process of responsible development of Artificial Intelligence, a part in which humanities scholars have a significant role to play.

Notes and References

[1] Avin, S., R. Gruetzemacher and J. Fox. 'Exploring AI futures through role play', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (February 2020), pp. 8–14.

[2] Shoham, Y., R. Perrault, E. Brynjolfsson and J. Clark. *Artificial Intelligence Index — 2017 Annual Report* (2017). https://aiindex.org/2017-report.pdf

3 Turing, A.M. ‘Computing machinery and intelligence’, *Mind, 49* (1950): 433–60.

4 Legg, S. and M. Hutter. ‘Universal intelligence: A definition of machine intelligence‘, *Minds and Machines*, *17*(4) (2007): 391–444. https://doi.org/10.1007/s11023-007-9079-x

5 Minsky, M. *Society of Mind*. Simon and Schuster (1988); Moravec, H. *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press (1988).

6 Campbell, M., A. J. Hoane Jr and F. H. Hsu. ‘Deep blue‘, *Artificial Intelligence*, *134*(1–2) (2002): 57–83; Silver, D., J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez ... and Y. Chen. ‘Mastering the game of Go without human knowledge‘, *Nature*, *550*(7676) (2017): 354. https://doi.org/10.15368/theses.2018.47

7 Krizhevsky, A., I. Sutskever and G. E. Hinton. ‘Imagenet classification with deep convolutional neural networks’, *Advances in Neural Information Processing* Systems (2012), pp. 1097–1105. https://doi.org/10.1145/3065386

8 LeCun, Y., Y. Bengio and G. Hinton. ‘Deep learning’, *Nature*, *521*(7553) (2015): 436. https://doi.org/10.1038/nature14539

9 Lippmann, R. (1987). An introduction to computing with neural nets. *IEEE Assp magazine*, *4*(2), 4–22.

10 Amodei, D. & Hernandez, D. (2018) AI and Compute. Open AI blog. Retrieved from https://blog.openai.com/ai-and-compute/.

11 Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun. ‘Dermatologist-level classification of skin cancer with deep neural networks’, *Nature*, *542*(7639) (2017): 115. https://doi.org/10.1038/nature21056

12 Silver et al. (2017).

13 Lyrebird. *We Create the Most Realistic Artificial Voices in the World* (2018). https://lyrebird.ai/

14 Suwajanakorn, S., S. M. Seitz and I. Kemelmacher-Shlizerman. ‘Synthesizing Obama: Learning lip sync from audio’, *ACM Transactions on Graphics* (*TOG*), *36*(4) (2017): 95. https://doi.org/10.1145/3072959.3073640

15 Fraze, D. *Cyber Grand Challenge* (*CGC*) (2018). https://www.darpa.mil/program/cyber-grand-challenge

16 Zhang, C., H. Li, X. Wang and X. Yang. ‘Cross-scene crowd counting via deep convolutional neural networks’, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 833–41. https://doi.org/10.1109/cvpr.2015.7298684

17 Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, ... and H. Anderson. ‘The malicious use of artificial intelligence: Forecasting, prevention, and mitigation’, *arXiv preprint arXiv:1802.07228* (2018).

18 Cave, S. & Dihal, K. (2018). Ancient dreams of intelligent machines: 3,000 years of robots. *Nature*, *559*(7715), 473. https://doi.org/10.1038/d41586-018-05773-y

19 Kurzweil, R. *The Singularity Is Near*. Gerald Duckworth & Co (2010).

20 Mozur, P. ‘Inside China’s dystopian dreams: AI, shame and lots of cameras‘, *New York Times* (8th July 2018). Turchin, A. and D. Denkenberger. ‘Classification of global catastrophic risks connected with artificial intelligence‘, *AI & SOCIETY* (2018): 1–17. https://doi.org/10.1007/s00146-018-0845-5

21 Bostrom, N. *Superintelligence: Paths, Dangers, Strategies* (2014).

22 Endy, D. *Synthetic Biology — What Should We Be Vibrating About*? TEDxStanford (2014). https://www.youtube.com/watch?v=rf5tTe_i7aA

23 De Vany, A. *Hollywood Economics: How Extreme Uncertainty Shapes the Film Industry*. Routledge (2004).

24 Yudkowsky, E. 'Artificial intelligence as a positive and negative factor in global risk', *Global Catastrophic Risks*, *1*(303) (2008): 184. https://doi.org/10.1093/oso/9780198570509.003.0021

25 Kareiva, P. and V. Carranza. 'Existential risk due to ecosystem collapse: Nature strikes back', *Futures* (2018); Needham, D. and J. Weitzdörfer (eds.). *Extremes* (Vol. 31). Cambridge University Press (2019). https://doi.org/10.1016/j.futures.2018.01.001

26 Kakoudaki, D. *Anatomy of a Robot: Literature, Cinema, and the Cultural Work of Artificial People*. Rutgers University Press (2014).

27 Harris, J. '16 AI bots with human names', *Chatbots Life* (2017). https://chatbotslife.com/10-ai-bots-with-human-names-7efd7047be34

28 Shanahan, M. *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press (2010).

29 Rose, H. 'Science fiction's memory of the future', *Contested Futures: A Sociology of Prospective Techno-Science*. Ashgate (2000): 157–74.

30 Oleszczuk, A. 'Sad and rabid puppies: Politicization of the Hugo Award nomination procedure', *New Horizons in English Studies*, (2) (2017): 127. https://doi.org/10.17951/nh.2017.2.127

31 De Vany (2004).

32 Kirby, D. A. *Lab Coats in Hollywood: Science, Scientists, and Cinema*. MIT Press (2011).

33 Boden, M. A. *AI: Its Nature and Future*. Oxford University Press (2016); Domingos, P. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books (2015); Shanahan, M. *The Technological Singularity*. MIT Press (2015).

34 Fry, H. *Hello World: How to Be Human in the Age of the Machine*. Penguin (2018); Tegmark, M. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Knopf (2017).

35 Kurzweil (2010).

36 Barrett, A. M. and S. D. Baum. 'A model of pathways to artificial superintelligence catastrophe for risk and decision analysis', *Journal of Experimental & Theoretical Artificial Intelligence*, *29*(2) (2017): 397–414; Turchin and Denkenberger (2018). https://doi.org/10.1080/0952813x.2016.1186228

37 Brynjolfsson, E. and A. McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company (2014); Hanson, R. *The Age of Em: Work, Love, and Life When Robots Rule the Earth*. Oxford University Press (2016).

38 Hyacinth, B. T. *The Future of Leadership: Rise of Automation, Robotics and Artificial Intelligence*. MBA Caribbean Organisation (2017); Rouhianien, L. *Artificial Intelligence: 101 Things You Must Know Today About Our Future*. Createspace Independent Publishing Platform (2018).

39 Srnicek, N. and A. Williams. *Inventing the Future: Postcapitalism and a World Without Work*. Verso Books (2015); Bastani, A. *Fully Automated Luxury Communism: A Manifesto*. Verso (2018).

40 Bostrom (2014).

41 Armstrong, S. and K. Sotala. 'How we're predicting AI — or failing to', *Beyond Artificial Intelligence* . Springer (2015), pp. 11–29. https://doi.org/10.1007/978-3-319-09668-1_2; Beard, S., T. Rowe and J. Fox. 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards', *Futures*, *115* (2020): 102469. https://doi.org/10.1016/j.futures.2019.102469; Yudkowsky, E. 'There's no fire alarm for artificial general intelligence', *Machine Intelligence Research Institute* (2017). https://intelligence.org/2017/10/13/fire-alarm/

42 Baum, S. 'Superintelligence skepticism as a political tool', *Information*, *9*(9) (2018): 209.

43 Grace, K., J. Salvatier, A. Dafoe, B. Zhang and O. Evans. 'When will AI exceed human performance? Evidence from AI experts', *arXiv preprint arXiv:1705.08807* (2017); Müller, V. C. and N. Bostrom. 'Future progress in artificial intelligence: A survey of expert opinion', *Fundamental Issues of Artificial Intelligence*. Springer (2016), pp. 555–72.

44 The Royal Society. *Public Views of Machine Learning* (2017). https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf

45 Achen, C. H. and L. M. Bartels. *Democracy for Realists: Why Elections Do Not Produce Responsive Government* (Vol. 4). Princeton University Press (2017).

46 Grosz, B. J. and P. Stone. 'A century long commitment to assessing Artificial Intelligence and its impact on society', *arXiv preprint arXiv:1808.07899* (2018).

47 Brundage, Avin et al. (2018).

48 Collins, H. M. and R. Evans. 'The third wave of science studies: Studies of expertise and experience', *Social Studies of Science*, *32*(2) (2002): 235–96. https://doi.org/10.1177/0306312702032002003; Owens, S. 'Three thoughts on the Third Wave', *Critical policy studies*, *5*(3) (2011): 329–33. https://doi.org/10.1080/19460171.2011.606307

49 British Academy and The Royal Society. *The Impact of Artificial Intelligence on Work* (2018). https://royalsociety.org/~/media/policy/projects/ai-and-work/evidence-synthesis-the-impact-of-AI-on-work.PDF

50 Donnelly, C. A., I. Boyd, P. Campbell, C. Craig, P. Vallance, M. Walport ... and C. Wormald. 'Four principles to make evidence synthesis more useful for policy', *Nature*, *558*(7710) (2018): 361. https://doi.org/10.1038/d41586-018-05414-4; Sutherland, W. J. and C. F. Wordley. 'A fresh approach to evidence synthesis', *Nature*, *558* (2018): 364–66. https://doi.org/10.1038/d41586-018-05472-8

51 Felten, E. and T. Lyons. *The Administration's Report on the Future of Artificial Intelligence* (2016). https://obamawhitehouse.archives.gov/blog/2016/10/12/administrations-report-future-artificial-intelligence; House of Lords. 'AI in the UK: Ready, willing and able?', *House of Lords Select Committee on Artificial Intelligence Report of Session 2017–19* (2018).

52 Farmer, J. D. and F. Lafond. 'How predictable is technological progress?', *Research Policy*, *45*(3) (2016): 647–65. https://doi.org/10.2139/ssrn.2566810

53 Eckersley, P. and Y. Nasser et al. *EFF AI Progress Measurement Project* (2017). https://eff.org/ai/metrics

54 Amodei and Hernandez (2018).

55 Martínez-Plumed, F., S. Avin, M. Brundage, A. Dafoe, S. ÓhÉigeartaigh and J. Hernández-Orallo. ‘Accounting for the Neglected Dimensions of AI Progress‘, *arXiv preprint arXiv:1806.00610* (2018).

56 Benaich, N. and I. Hogarth. *The State of Artificial Intelligence in 2018: A Good Old-Fashioned Report* (2018). https://www.stateof.ai/; Shoham et al. (2017).

57 Owen, R., P. Macnaghten and J. Stilgoe. ‘Responsible research and innovation: From science in society to science for society, with society’, *Science and Public Policy*, *39*(6) (2012): 751–60. https://doi.org/10.4324/9781003074960-11

58 Stilgoe, J. ‘Machine learning, social learning and the governance of self-driving cars’, *Social Studies of Science*, *48*(1) (2018): 25–56. https://doi.org/10.2139/ssrn.2937316

59 Jungk, R. and N. Müllert. *Future Workshops: How to Create Desirable Futures*. Institute for Social Inventions (1987).

60 Nikolova, B. ‘The rise and promise of participatory foresight‘, *European Journal of Futures Research*, *2*(1) (2014): 33. https://doi.org/10.1007/s40309-013-0033-2; Oliverio, V. *Participatory Foresight*. Centre for Strategic Futures (2017). https://www.csf.gov.sg/our-work/Publications/Publication/Index/participatory-foresight

61 Amer, M., T. U. Daim and A. Jetter. ‘A review of scenario planning’, *Futures*, *46* (2013): 23–40. https://doi.org/10.1016/j.futures.2012.10.003

62 Perla, P. *The Art of Wargaming: A Guide for Professionals and Hobbyists* (Vol. 89028818). Naval Institute Press (1990).

63 Bryant, J. *The Six Dilemmas of Collaboration: Inter-Organisational Relationships as Drama*. Wiley (2002).

64 Cohen, T., J. Stilgoe and C. Cavoli. ‘Reframing the governance of automotive automation: Insights from UK stakeholder workshops’, *Journal of Responsible Innovation* (2018): 1–23. https://doi.org/10.1080/23299460.2018.1495030

65 Kitcher, P. *Science in a Democratic Society*. Prometheus Books (2011).

66 Deconstructeam. *The Red Strings Club* [PC game]. Devolver Digital (2018).

67 Lantz, F. (2017) Universal Paperclips [online video game]. Retrieved from http://www.decisionproblem.com/paperclips/

68 Bostrom (2014).

69 Paradox Interactive. *Stellaris* [video game] (2016).

70 Shapira, S. and S. Avin. *Superintelligence* [video game mod] (2017). https://steamcommunity.com/sharedfiles/filedetails/?id=1215263272

71 Firaxis Games. *Sid Meier's Civilization V* [PC game] (2010).

72 Fantasy Flight Games. *End of the World: Revolt of the Machines* [roleplaying game] (2016).

73 Ligan, F. *Mutant: Mechatron* [roleplaying game] (2017).

9. Accumulating Evidence Using Crowdsourcing and Machine Learning: A Living Bibliography About Existential Risk and Global Catastrophic Risk

Gorm E. Shackelford, Luke Kemp, Catherine Rhodes, Lalitha Sundaram, Seán S. ÓhÉigeartaigh, SJ Beard, Haydn Belfield, Julius Weitzdörfer, Shahar Avin, Dag Sørebø, Elliot M. Jones, John B. Hume, David Price, David Pyle, Daniel Hurt, Theodore Stone, Harry Watkins, Lydia Collas, Bryony C. Cade, Thomas Frederick Johnson, Zachary Freitas-Groff, David Denkenberger, Michael Levot and William J. Sutherland

Highlights:

- This chapter presents a semi-automated process for systematically reviewing the relevance of academic research to the study of existential risk to provide an evidence base for policy and risk analysis. Despite its recent emergence and neglected status, the growth and interdisciplinary scope of Existential Risk Studies means that an overwhelming volume of relevant research has already been published.
- In a systematic review, one of many time-consuming tasks is to read the titles and abstracts of research publications, to see if they meet the inclusion criteria. This chapter shows how this task can be shared between multiple people (using crowdsourcing) and partially automated (using machine learning).

https://doi.org/10.11647/OBP.0360.09

- The authors used these methods to create The Existential Risk Research Assessment (TERRA), which is a living bibliography of relevant publications that gets updated each month and is freely available at terra.cser.ac.uk.
- The chapter presents the results from the first 10 months of TERRA, in which 10,001 abstracts were screened by 51 participants, highlighting the potential and challenges of this approach and recommending that, for now, semi-automated tools like this should be used in tandem with manually curated bibliographies.
- The authors note that a number of challenges remain, including trade-offs between recall (inclusion of all relevant research) and accuracy (exclusion of irrelevant research), different levels and domains of expertise among assessors, and the incomplete assessment of training data. However, they suggest that "collaborative and cumulative methods" such as these will need to be used in systematic reviews as the volume of research increases.

This chapter was originally published in *Futures* in 2020 but TERRA continues to be maintained and updated by CSER. If you would like to help to continue training our algorithm or sign up for monthly updates of new research, please go to terra.cser.ac.uk. The TERRA database was used as part of the research process for Chapter 23 of this volume, whilst the utilisation of semi-automated tools is discussed further in Chapter 7.

> In the past, censorship worked by blocking the flow of information. In the twenty-first century censorship works by flooding people with irrelevant information. [...] Today having power means knowing what to ignore.
>
> — Yuval Noah Harari, *Homo Deus* (p. 462)

1. Introduction

An overwhelming volume of research has been published in recent years. There is now a deep division (called the "synthesis gap") between research that has been published and research that has been systematically reviewed, synthesised, and used for decision making.[1] We need new methods of quickly and efficiently finding relevant research,[2] and we need these methods to be rigorous, transparent, and inclusive,[3] to minimise bias in the decisions that are based on this research. Bad decisions can mean death or extinction in some fields (e.g. medicine or wildlife conservation),[4] and it may be vitally important to develop more efficient methods of reviewing research and using it for evidence-based decision-making in these fields.[5] The need for more efficient methods of reviewing research could be even more important when considering existential risks and Global Catastrophic Risks, because the consequences of bad decisions could be disastrous, and yet decisions will need to be made in the near future about which interventions should be used to reduce these risks.[6]

Research on nuclear weapons, published in the early years of the Cold War, has been seen as some of the earliest research on existential risks or Global Catastrophic Risks.[7] However, an integrated field of research on existential risks and Global Catastrophic Risks as special classes of risk has only recently emerged.[8] We will refer to these risks collectively as "existential risks" or "x-risks" hereafter. Many research centres in this field have only recently become established, such as the Future of Humanity Institute (FHI) at the University of Oxford in 2005, the Global Catastrophic Risk Research Institute (GCRI) in 2011, the Centre for the Study of Existential Risk (CSER) at the University of Cambridge in 2012, and programmes at the universities of Copenhagen, Gothenburg (Chalmers), and Warwick. However, an overwhelming volume of research on existential risk already exists, because research from well-established fields, such as Artificial Intelligence, biosecurity, climate science, ecology, and philosophy, is also relevant to the integrated study of existential risks. Thus, the volume of relevant research on existential risks is perhaps even more overwhelming than it is in many other fields.

To support evidence-based decision-making about existential risks, this research should ideally be systematically reviewed. A systematic review is an effort to review all evidence on a research question (e.g. "What are the effects on this drug on this disease?" or, in the context of existential risk, "What are the likely impacts of this risk on human civilization?"), while minimising bias in the evidence base.[9] It is often assumed that the best evidence for an evidence-based decision will come from systematic reviews,[10] but there are other methods of reviewing research (e.g. "subject-wide evidence synthesis"), which could also be useful for a field as broad as existential risk. In the context of this publication, we refer to any information that could be used to support decision-making as "evidence" (e.g. not only scientific data but also philosophical arguments), and we refer to "systematic reviews" of this evidence, but our methods are also relevant to other forms of evidence synthesis.

We show how an overwhelming volume of research publications can be screened for inclusion in a systematic review, using crowdsourcing and machine learning, and how the relevant publications can be accumulated in an open-access database that can be reused repeatedly. The "synthesis gap" is a problem in many fields, and a solution to this problem could have broad applications in other fields. However, the methods we use here are only a partial solution to this problem. Screening publications for inclusion is only one of many tasks in a systematic review, and much more research will be needed before evidence can be extracted from these publications, and before the synthesis gap can be closed.

Machine learning can be used to predict the relevance of publications to a systematic review, using "text mining".[11] Based on a "training set" of publications that have been labelled as "relevant" or "irrelevant" by humans, a machine-learning classifier can be trained to predict which publications are relevant, using the text in their titles and/or abstracts. The accuracy of the classifier can be tested using a "test set" of publications that have also been labelled by humans, and the relevance of a new set of publications that have not yet been screened by humans can then be predicted by the classifier.[12] By using text mining, the human workload can be reduced by 30–70% when screening publications for systematic reviews.[13]

Crowdsourcing can also be used when screening publications,[14] and by sharing the workload between multiple people, the time and/or money it takes can be reduced. For example, the cost was reduced by 88% in a test of using crowdsourcing to screen publications.[15] If the evidence base can be updated and reused (which we refer to as "evidence accumulation"), then crowdsourcing can also save time and/or money by sharing the workload between the past, present, and future. Crowdsourcing is used by Cochrane (the collaboration for systematic reviews in medicine that has set the standard for other fields of research), in the form of the "Cochrane Crowd" (http://crowd.cochrane.org).

Crowdsourcing is also used in futures studies, as a method of horizon scanning for emerging threats.[16]

For crowdsourcing and evidence accumulation to work well over time, the evidence that we are beginning to accumulate now will need to be relevant to the research and policy questions that are asked in the future. Two related methods of accumulating evidence, which are likely to be relevant to future research and policy questions, are "systematic mapping" and "subject-wide evidence synthesis",[17] in which a wide-ranging search strategy is used to find publications that are relevant to a whole subject (e.g. existential risk), rather than using a narrower search strategy that cannot contribute to future research on related topics within that subject. Publications from a wider search can later be classified into narrower topics, and the systematic map can be updated and reused to answer narrower questions in the future, without needing to begin a new search for each narrower topic. Our approach follows the principles of subject-wide evidence synthesis, using crowdsourcing, machine learning, and evidence accumulation in an open-access online database to create a bibliography of publications about existential risk. We called this process "The Existential Risk Research Assessment" (TERRA).

There are already several "conventional" bibliographies of existential risk research (i.e. bibliographies without crowdsourcing or machine learning), including the "Global Challenges Bibliography" in Appendix 1 of *Global Challenges: 12 Risks that Threaten Human Civilization*,[18] the "Bibliography of Collapse" (http://www.collapsologie.fr), and bibliographies from research centres such as FHI (https://www.fhi.ox.ac.uk/publications/) and GCRI (https://gcrinstitute.org/publications/). Although these bibliographies are useful resources for the research community, they are not based on transparent search strategies with clearly stated inclusion criteria, which are vital principles for systematic reviews,[19] and which would make these bibliographies more useful for future research. In contrast, our approach is based on four principles that are recommended for research synthesis:[20] "transparency" (clearly stating our search strategy and inclusion criteria), "rigorousness" (repeating the process with multiple participants, and minimising bias by using a broad search strategy, but not yet being truly comprehensive), "inclusiveness" (including the research community as participants in the screening process), and "accessibility" (being freely available online).

2. Methods

2.1 Summary of the methods

We used keywords to search for publications about existential risk. Based on the titles and/or abstracts of these publications, we labelled each publication as "relevant" or "irrelevant" to existential risk. A bibliography of "relevant" publications is freely available for downloading as CSV and RIS files from terra.cser.ac.uk. We used these labelled publications to train a machine-learning classifier. We then set up an automated and regularly scheduled search for new publications, using the same keywords. The machine-learning classifier predicts the relevance of the new publications, and the list of the new publications that it predicts to be relevant are emailed to the participants, but these publications are not added to the bibliography until they have been assessed by at least one person.

2.2 Search strategy

Our search strategy was based on the "Global Challenges Bibliography" in Appendix 1 of *Global Challenges: 12 Risks that Threaten Human Civilization*,[21] which included publications up to 2013, and which was the most systematically collected bibliography about existential risks of which we were aware. We used the keywords that were used for the Global Challenges Bibliography to search the titles, abstracts, keywords, and references of publications in *Scopus* in 2017. We then compared our search results with the publications in the Global Challenges Bibliography. If a publication in the Global Challenges Bibliography was not in the search results, but it was in *Scopus*, then we added keywords that would find this publication (unless there were no keywords that seemed specific enough to existential risk to justify their use). Using this extended set of keywords, we then searched *Scopus* again, and we continue to search it regularly for new publications (see below for search terms). We acknowledge that this is not the only possible search strategy, and *Scopus* is not the only database of publications, but it was the only database to which we had programmatic access through an API (Application Programming Interface), which we needed to automate the monthly searches. These limitations should be considered when using our bibliography as part of a systematic review. Nevertheless, our bibliography represents a more systematic and comprehensive approach to mapping the literature on existential risk than any other approach of which we are aware, and thus it represents significant progress.

2.3 Search terms

Title-Abstract-Keywords: "catastrophic risk" OR "existential risk" OR "existential catastrophe" OR "global catastrophe" OR "human extinction" OR "infinite risk" OR "xrisk" OR "x-risk" OR apocalypse OR doomsday OR doom OR "extinction of human" OR "extinction of the human" OR "end of the world" OR "world's end" OR "world ending" OR "end of civilization" OR "collapse of civilization" OR "survival of civilization" OR "survival of humanity" OR "human survival" OR "survival of human" OR "survival of the human" OR "global collapse" OR "historical collapse" OR "catastrophic collapse" OR "global disaster" OR "existential threat" OR "catastrophic harm"

References: "catastrophic risk" OR "existential risk" OR "existential catastrophe" OR "global catastrophe" OR "human extinction" OR "infinite risk" OR "xrisk" OR "x-risk"

2.4 Inclusion criteria

We used the following inclusion criteria as guidelines for assessing publications as "relevant" or "irrelevant" to existential risk or Global Catastrophic Risk (copied from the website):

For the purpose of this assessment, a risk is "catastrophic" if it causes at least 10 million deaths (approximately) and a risk is "existential" if it causes the extinction of the human species or the collapse of human civilisation.[22] Publications that are relevant do not need to include the exact phrase "existential risk" or "Global Catastrophic Risk" but they should be about a risk that is *global* and *catastrophic* in scale.

Publications that are relevant should *explicitly* be about the possibility, probability, impact, or management of existential or global catastrophic risks, as opposed to other aspects of these risks that are only implicitly relevant. For example, a publication about the probability of an asteroid impact that could kill all humans should be included, whereas a publication about some other aspect of an asteroid impact (e.g. the geological evidence of an asteroid impact in the past) should not be included. A publication about climate change should be included only if it is about *global catastrophic* climate change. Likewise, a publication about insurance against catastrophic risk should be included only if it is about *Global* Catastrophic Risk (and loss of life, as opposed to financial loss), and a publication about disaster management should be included only if it is about a *global* disaster (as opposed to a global response to a local or regional disaster).

Alternatively, a more common-sense criterion is to ask whether or not a publication is really *about* existential risk or *about* Global Catastrophic Risk, rather than something that is only tangentially related to such a risk. Many publications seem to make passing reference to things that are allegedly essential to human survival without actually discussing them as such.

Relevant publications should include at least one criterion from the following list.

- Discussion of existential risk or Global Catastrophic Risk *per se* (explicit, not implicit)
- Assessment of such a risk (e.g. the probability or impact of nuclear winter in the event of nuclear war)
- Discussion of a strategy for managing such a risk (e.g. strategic food reserves to mitigate the risk of human extinction from catastrophes that destroy crops)
- Comparison of these risks (e.g. the relative risk of human extinction from asteroid impact compared to Artificial Intelligence)
- Philosophical discussion that is relevant to these risks (e.g. the "value" of the future lives that would be saved by preventing the extinction of the human species)

Publications about artistic, fictional, or religious works should not be included.

2.5 Crowdsourcing

TERRA is based at the Centre for the Study of Existential Risk (CSER) at the University of Cambridge. To recruit participants from outside of CSER, we promoted TERRA on social media (Facebook and Twitter), on the CSER website (www.cser.ac.uk), and in a workshop at the Cambridge Conference on Catastrophic Risk (17–18 April 2018). Participation was open to anyone. Anyone who assessed at least 500 publications as of 31 August 2018 was invited to be a co-author of this publication.

TERRA is a web application that is hosted at terra.cser.ac.uk and is based on the *Django* framework for *Python* (www.djangoproject.com). When using the web app, each participant is shown titles and abstract from our search results (in a random order, to minimise bias) and is asked to assess the relevance of each publication based on the inclusion criteria (see above). Each participant is also asked to assess the relevance of each publication to each specific class of risk (such as "Artificial Intelligence" or "biotechnology"). We developed a system of classifying existential risks (Figure 1) for the purposes of classifying publications for TERRA, but other classification systems could be used for other purposes, such as integrated risk assessment.[23]

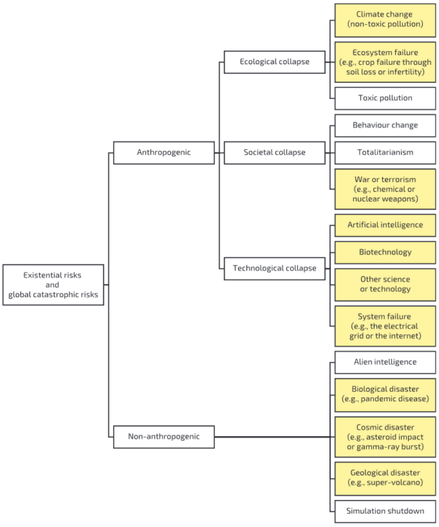


Fig. 1: The classification of existential risks and global catastrophic risks that we developed for The Existential Risk Research Assessment (TERRA). The classes that were used to tag publications are highlighted in yellow.

Different people are likely to have made different decisions about the relevance of each publication, not only because existential risk is an emerging field with blurry boundaries, but also because different people have different disciplinary backgrounds, personal worldviews, subjective biases, and so on. Therefore, to test the consistency of these decisions about the relevance of each publication, we calculated the "agreement" between the people who assessed each publication. For example, if a publication was assessed as "relevant" by either 0% or 100% of the people that assessed it, then there was 100% agreement between these people. If a publication was assessed as "relevant" by 50% of the people that assessed it, then there was 50% agreement. We plotted agreement by class of risk, and we used Wilcox tests in *R* to test whether agreement about publications with a specified class of risk was different from agreement about publications without a specified class (i.e. publications about generalised risks). We also plotted the number of publications and the number of "relevant" publications over time, to test the rate of increase (see "Results and Discussion").

2.6 Machine learning

We used an artificial neural network, implemented in the *TensorFlow* library for *Python* (www.tensorflow.org), to predict the relevance of publications that had not yet been assessed by humans, based on the abstracts of publications that had been assessed (labelled as "relevant" or "irrelevant"). First, we excluded the publications that had been assessed but did not have abstracts (because we wanted to use the abstract to make the predictions). Second, we randomly split the publications that had been assessed into a training set (80% of publications) and a test set (20% of publications). Third, we used the first 200 words of each abstract in the training set (labelled as "relevant" or "irrelevant") as the inputs into the neural network (200 was the average number of words in these abstracts), and we used a "convolution" layer in the network to encode each of these words as a vector of numbers ("word embedding"), based on its relationship to the other words in the abstract. Fourth, we passed these word embeddings to a fully connected layer in the network. When the network was trained, we used it to predict the probability that each publication in the test set was relevant. These methods were based on methods described by Géron.[24]

We then generated three different models, by setting three different probability thresholds to control the unavoidable trade-off between "precision" and "recall".[25] Precision is the percentage of publications that were predicted to be relevant by the machine that are "truly" relevant. Recall is the percentage of truly relevant publications that were correctly predicted to be relevant by the machine. We generated "low-recall", "medium-recall", and "high-recall" models that aim for 50%, 75%, and 95% recall, respectively. The trade-off is that the models with higher recall have lower precision, and so they save less time in finding truly relevant publications, but they are less likely to miss truly relevant publications. We used these models to predict the relevance of publications that had not yet been assessed by humans. Users can choose the model that makes the most sense for their use-cases, and these trade-offs are explained on the website.

3. Results and Discussion

3.1 Crowdsourcing

By 31 August 2018, a total of 12,635 publications had been included in the database. A total of 51 people had assessed at least one publication, and 19 of these people had assessed at least 500 publications, including eight people from CSER (the first eight authors of this publication). Many of the other participants were previously unknown to CSER, and so this project is helping us to recruit new participants to our research network. A total of 10,001 publications were assessed by at least one person (79% of publications in the database), and 2,313 of these 10,001 publications (23%) were assessed as "relevant" by at least one person.

Of these 10,001 publications, 5,961 were assessed by at least two people (47% of the publications in the database), and we analysed the agreement between different people for these publications. Of these 5,961 publications, 1,722 (29%) were assessed as "relevant" by at least one person. For each publication that was assessed as "relevant" by one person, there was approximately one other person who assessed that same publication as "irrelevant" (there was 56% agreement between assessors). However, there was higher agreement overall, when including publications that everyone assessed as “irrelevant” (87% agreement). Thus, there was higher agreement about what to exclude than what to include. Only 628 of these 5,961 publications (11%) were assessed as “relevant” by at least two people. Unsurprisingly, this suggests that the literature about existential risks and global catastrophic risks is difficult to define (because it is an emerging and wide-ranging field). In the future, when more people have assessed each publication, we hope to be able to use the data on agreement for more sophisticated analyses,[26] but at present we use it only to rank the publications in the bibliography, first by relevance (the number of “relevant” assessments minus the number of “irrelevant” assessments) and second (within each level of relevance) by the total number of assessments.

The highest-ranked publications are inevitably among those that have been assessed the most, but the lowest-ranked publications are also inevitably among those that have been assessed the most. We think this is sensible, because we have the most information about these publications, and so we have the most confidence in whether they are seen as relevant or irrelevant. However, a systematic reviewer would presumably need to consider all publications that at least one person had assessed as relevant, rather than considering only the highest-ranked publications (and indeed the downloadable bibliography includes all publications that at least one person assessed as relevant). The reason that some publications are assessed more than others is partly by chance (participants are shown titles and abstracts in a random order) and partly by choice (participants are also sent a monthly email, with recent publications that the machine-learning model has predicted to be relevant, and they are asked to assess these publications, and they are also asked not to assess a publication if they are uncertain about its relevance). Thus, the highest-ranked and lowest-ranked publications are more likely to be recent publications (published in or after November 2017, when the monthly email began to be sent), because recent publications are more likely to be assessed by multiple people, and they are also likely to be publications about which people had greater certainty. For this reason, the ranking should only be seen as a starting point for future studies. For example, it would be possible to download the bibliography and reorder it by average relevance, or simply to read through all relevant titles in a random order.

Of the 1,722 publications for which we analysed agreement, the publications that were also assessed as "relevant" to a specified class of risk (Figure 2) often had higher agreement than the mean agreement for all publications (Figure 3). Publications about Artificial Intelligence, biological disaster, biotechnology, climate change, or cosmic disaster had significantly higher agreement than the mean, and publications about biotechnology had the highest agreement (74%), but publications about ecosystem failure, geological disaster, other science or technology, system failure, and war or terrorism had agreement that was not significantly different from the mean. This suggest that some risks could be more definitive of existential risk as a field of research. If so, we should beware of marginalising these other less distinctive risks in our thinking about existential risk as a field. However, it is also possible that these risks could be more distinctive because they are bigger risks. Moreover, it is possible that these patterns could be caused by sampling bias, since the participants were not randomly sampled, and they should not necessarily be seen as representative of the global existential risk research community.

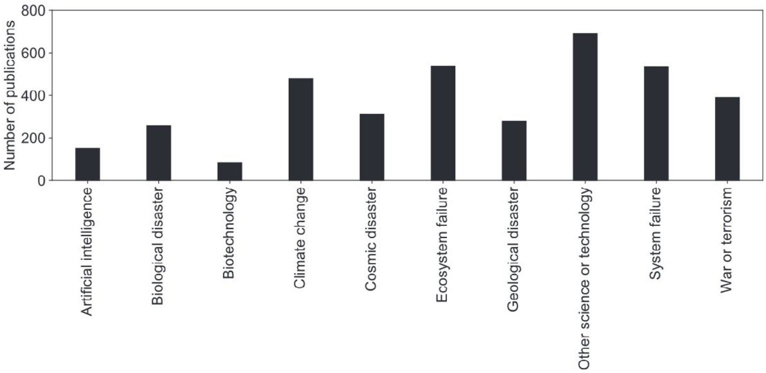


Fig. 2: Number of publications that were indexed in *Scopus*, found using our search strategy, and assessed as "relevant" to at least one specified class of existential risk or Global Catastrophic Risk by at least one person as of 31 August 2018.

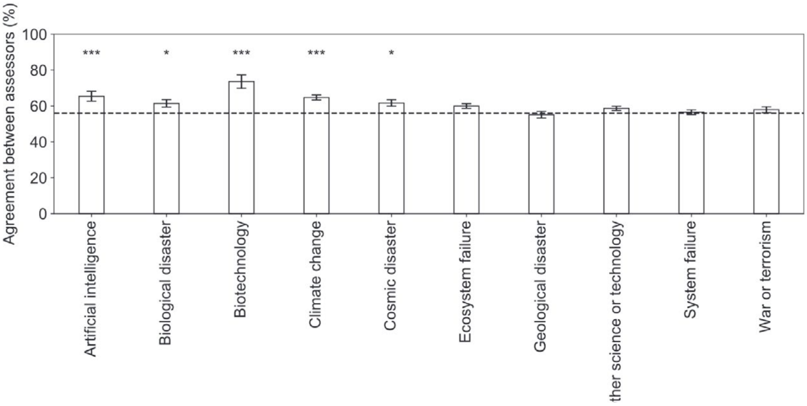


Fig. 3: Agreement between assessors as a function of the class of risk. The dotted line shows the mean agreement (56%) for all publications that were assessed as "relevant" by at least one person. The error bars show one standard error above and below each mean. Significant differences from the mean for all publications (the dotted line) are shown with asterisks ("*": $P < 0.05$; "***": $P < 0.001$; P-values from Wilcox tests).

The number of publications that have been found by our search strategy is increasing over time at an exponential rate (Figure 4). This is an unsurprising but concerning trend that has also been reported in other fields, and indeed this trend is the motivation for using these new methods of evidence synthesis.[27] However, what is surprising and may be even more concerning is that the number of "relevant" publications, as a proportion of the total number of publications, is decreasing over time (Figure 4). In other words, to find one "relevant" publication, we now need to review more publications than we did in the past. Another surprising finding is that there appears to have been a rapid increase in the number of publications after the year 2000, followed by a rapid decrease after 2010.

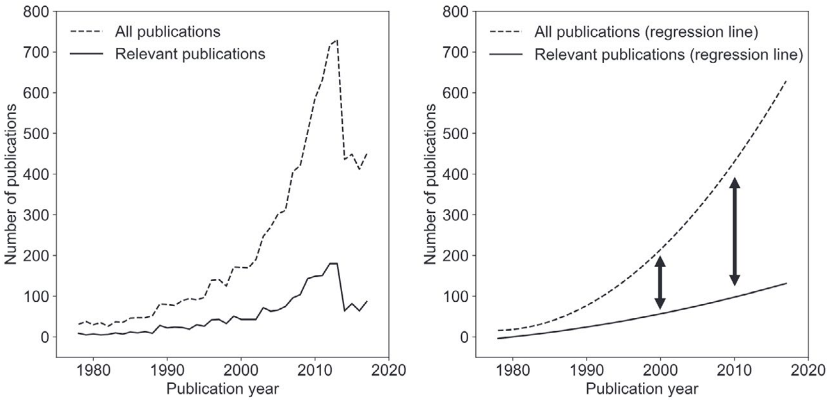


Fig. 4: Number of publications in the 40 years from 1978–2017 that were indexed in *Scopus*, found using our search strategy, and assessed as "relevant" by at least one person as of 31 August 2018 (when only 79% of publications had been assessed, and thus this is an underestimate of the total number, but a reasonable estimate of the trend, since publications were assessed in a random order). The trend lines show the widening gap between all publications and relevant publications over time, and their equations are $y = -1555.64x + 0.393332x^2 + 1538163$ for all publications and $y = -166.197714x + 0.04247x^2 + 162574$ for relevant publications (for the years 1978–2017).

3.2 Machine learning

We were pleased to see that the search strategy had successfully found, and the neural network had correctly predicted, the relevance of several recent publications that we already knew to be relevant to existential risk,[28] but that is only anecdotal evidence. Based on the test set of publications in August 2018, the low-recall bibliography had a precision of about 50% and a recall of about 50%, the medium-recall bibliography had a precision of about 33% and a recall of about 75%, and the high-recall bibliography had a precision of about 24% and a recall of about 96% (Table 1).

Table 1: Trade-off between precision and recall in the three machine-learning models as of 31 August 2018. "Precision" is the percentage of publications that were predicted to be relevant that are truly relevant. "Recall" is the percentage of truly relevant publications that were correctly predicted to be relevant. Precision and recall are estimates (based on the test set, and thus not necessarily representative of the prediction set). "Positives" is the number of unassessed publications that was predicted to be relevant. "True positives" = positives * precision (and thus it is also an estimate, because it is based on the estimate of precision).

Model	**Recall**	**Precision**	**Positives**	**True positives**
"High recall"	0.9589	0.2422	1258	305
"Medium recall"	0.7534	0.3343	696	233
"Low recall"	0.5000	0.5034	243	122

Of the 2,758 publications that had not yet been assessed by humans on 28 August 2018, when the neural network was retrained, perhaps 303 were truly relevant (11% of 2,758 publications, based on the 11% of publications that had been assessed as relevant by more than one person, as reported above, but this is only an estimate). When assessed by the neural network, 1,258 publications were included in the high-recall bibliography, and perhaps 305 were truly relevant (24% precision, based on the test set, but precision and recall are only estimates for the prediction set), which is similar to our estimate of 303 truly relevant publications. Thus, the high-recall bibliography would save time, because only 1,258 of 2,758 publications (46%) would need to be assessed by humans, and only 4% of truly relevant publications would have been excluded (96% recall). The amount of time that this would save would depend on how much time it would have taken to assess the machine-excluded publications (and many irrelevant publications are quick for humans to exclude). The low-recall bibliography would save more time, because only 243 of 2,758 publications (9%) would need to be assessed by humans, but 50% of truly relevant publications would have been excluded (50% recall).

Thus, the neural network seems to work well as a "recommendation engine" (automatically recommending the most relevant publications by email), and it could possibly also be used as an acceptable substitute for manual screening in systematic reviews, if 100% recall is not critical. However, in the short term, machine learning seems most useful for rapid evidence synthesis, in which timeliness is more important than comprehensiveness.[29] In the long term, if crowdsourcing and evidence accumulation can be used to share the workload between multiple people and multiple years, then machine learning seems less useful, unless there is an improvement in both precision and recall at the same time (using a larger or better training set or a better algorithm).

3.3 Limitations of these methods

TERRA has several limitations that should be considered before it is used in systematic reviews. One limitation is that participants have different levels of expertise in existential risk, and different views about the relevance of publications. However, participants were asked not to assess a publication if they were uncertain about its relevance, or else to be overly inclusive if they were ambivalent, and so TERRA is not likely to exclude relevant publications because of a lack of expertise. Disagreements between participants are interesting in themselves, and they could be an insight into existential risk as a research field. However, differences in expertise and differences of opinion could lead to different types of disagreement, and these different types of disagreement should be explored in the future. TERRA also offers an opportunity to learn more about existential risk by participating in the evidence assessment, and thus the expertise of participants could also increase over time.

Another limitation of TERRA is that 21% of the publications in the search results have not yet been assessed by anyone, and many publications have been assessed by only one person. Thus, the relevance of some publications is inconclusive. Another limitation is that only one database is being searched (*Scopus*). This will hopefully be resolved when other databases (such as *Web of Science*) offer free and easy access through an API. At present, *Scopus* is primarily focused on academic journal articles, and it does not include many books and popular texts on existential risk, such as *Our Final Century*.[30] Thus, this bibliography should be used in conjunction with other bibliographies, such as the Global Challenges Bibliography,[31] for increased comprehensiveness.

3.4 Towards a Doomsday Database

TERRA is helping to build a network of x-risk researchers, who in time could collaborate on systematically mapping and reviewing x-risk research. We can envision a "Doomsday Database" that would include all of the available evidence on the probabilities and impacts of each class of risk, based on data extracted from the literature. This evidence base could be used to compare different classes of risk and prioritise the risks with the highest probabilities and/or impacts, as part of the "integrated assessment" of risks.[32] For example, risks that have impacts on similar "critical systems" (e.g., food systems or security systems) or have similar "spread mechanisms" (e.g., biological or digital replicators) could be prioritised for simultaneous management.[33]

It is difficult to see how we could get from "here" (a crowdsourced bibliography) to "there" (a subject-wide database of probabilities and impacts). It was suggested in the Global Challenges Bibliography that the literature on some risks is "too voluminous to catalogue" (e.g. climate change), and this is one reason that we limited ourselves to a search for publications about existential risk in general. Although it was once suggested that there were fewer publications on "human extinction" than on "dung beetles",[34] our subject-wide view of the literature on existential risks shows that indeed it is voluminous and it is increasing at an exponential rate.

However, examples of such subject-wide databases exist. For example, the Conservation Evidence project (www.conservationevidence.com) is making progress towards a subject-wide database for the effectiveness of all conservation actions.[35] It is only by imagining the possibility of such a database for existential risks that we might make progress towards it. Moreover, the further development of crowdsourcing and machine learning may make it easier to imagine this scale of evidence synthesis in the near future. If it proves to be impossible to synthesise the evidence across all existential risks, on a subject-wide scale, then the methods that we have developed for TERRA could be used to search for publications about narrower topics (e.g. Artificial Intelligence), and a database could be developed for each of these topics.

An accessible, inclusive, rigorous, and transparent database could be especially useful for the governance of existential risk, considering the catastrophic consequences that policy failures could have (e.g. human extinction), and also considering the probability that the beneficial uses of new technologies will be promoted more than their harmful uses (for "dual-use technologies" such as genetic engineering and molecular nanotechnology). As well as evidence in a narrow sense, this database could also provide information about our collective understanding of existential risk. This would be evidence in broad sense (a "knowledge base"), and it could be used to support philosophical arguments about the definition of existential risk, and also to communicate existential risk to the public.

4. Conclusions

TERRA produces a regularly updated bibliography about existential risks. By including a wide range of participants (as "stakeholders" in existential risk research), by comparing their assessments, and by clearly reporting its methods, TERRA follows the recommendations that evidence synthesis should be accessible, inclusive, robust, and transparent.[36] As well as these strengths, TERRA also has limitations that should be considered before it is used in systematic reviews. These limitations are not insurmountable, and readers are invited to participate in TERRA and contribute to a bigger and better bibliography in the future.

5. Acknowledgements

This project was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation. Several of the authors were also supported by the David and Claudia Harding Foundation. We thank our funders, and we also thank Stuart Armstrong, Seth Baum, Sebastian Farquhar, Nancy Ockendon, Martin Rees, Jens Steffensen, Emile Torres, and all of the participants in TERRA.

Notes and References

[1] Westgate, Martin J. et al. 'Software support for environmental evidence synthesis', *Nature Ecology & Evolution, 2*(4) (1 April 2018): 588–90. https://doi.org/10.1038/s41559-018-0502-x

[2] e.g., Wallace, Byron C. et al. 'Modernizing the systematic review process to inform comparative effectiveness: Tools and methods', *Journal of Comparative Effectiveness Research, 2*(3) (1 May 2013): 273–82. https://doi.org/10.2217/cer.13.17; O'Mara-Eves, Alison et al. 'Using text mining for study identification in systematic reviews: A systematic review of current approaches', *Systematic Reviews, 4*(1) (14 January 2015): 5. https://doi.org/10.1186/2046-4053-4-5; Westgate et al. (2018).

[3] Donnelly, Christl A. et al. 'Four principles for synthesizing evidence', *Nature, 558*(7710) (2018): 361. https://doi.org/10.1038/d41586-018-05414-4

[4] Sutherland, William J. et al. 'The need for evidence-based conservation', *Trends in Ecology & Evolution, 19*(6) (2004): 305–8. https://doi.org/10.1016/j.tree.2004.03.018

[5] Sutherland, William J. and Claire F. R. Wordley. 'A fresh approach to evidence synthesis', *Nature, 558*(7710) (2018): 364. https://doi.org/10.1038/d41586-018-05472-8; Shackelford, Gorm E. et al. 'Evidence synthesis as the basis for decision analysis: A method of selecting the best agricultural practices for multiple ecosystem services', *Frontiers in Sustainable Food Systems, 3* (2019): 83. https://doi.org/10.3389/fsufs.2019.00083

[6] Wilson, Grant. 'Minimizing global catastrophic and existential risks from emerging technologies through international law', *Virginia Environmental Law Journal, 31*(2) (2013): 307–64; Farquhar, Sebastian et al. *Existential Risk: Diplomacy and Governance*. Global Priorities Project (2017); Brundage, Miles et al. *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* [report] (30 April 2018). https://doi.org/10.17863/CAM.22520

[7] e.g. Konopinski, E. J., C. Marvin and Edward Teller. 'Ignition of the atmosphere with nuclear bombs', *Report LA-602. Los Alamos, NM: Los Alamos Laboratory* (1946), cited in Dennis Pamlin and Stuart Armstrong, *Global Challenges: 12 Risks That Threaten Human Civilization*. Global Challenges Foundation (2015).

[8] e.g. Bostrom, Nick. 'Existential risks: Analyzing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9* (2002); Rees, Martin. *Our Final Century*. William Heinemann (2003); Bostrom, Nick and Milan M. Ćirković. *Global Catastrophic Risks*. Oxford University Press (2008); Bostrom, Nick. 'Existential risk prevention as global priority', *Global Policy, 4*(1) (27 March 2013): 15–31. https://doi.org/10.1111/1758-5899.12002; Avin, Shahar et al. 'Classifying global catastrophic risks', *Futures, 102* (2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001; please note that these citations should not be seen as a comprehensive chronology of the field.

[9] e.g. see Haddaway, N. R. et al. 'Making literature reviews more reliable through application of lessons from systematic reviews', *Conservation Biology, 29*(6) (1 June 2015): 1596–1605. https://doi.org/10.1111/cobi.12541 for some methods used in systematic reviews compared to non-systematic reviews.

10 e.g. Donnelly et al. (2018).

11 O'Mara-Eves et al. (2015).

12 Wallace et al. (2013); Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media, Inc. (2017).

13 O'Mara-Eves et al. (2015).

14 Brown, Andrew W. and David B. Allison. 'Using crowdsourcing to evaluate published scientific literature: Methods and example', *PLOS ONE, 9*(7) (2 July 2014): e100647. https://doi.org/10.1371/journal.pone.0100647; Mortensen, Michael L. et al. 'An exploration of crowdsourcing citation screening for systematic reviews', *Research Synthesis Methods, 8*(3) (4 July 2017): 366–86. https://doi.org/10.1002/jrsm.1252; Krivosheev, Evgeny, Fabio Casati, and Boualem Benatallah. 'Crowd-based multi-predicate screening of papers in literature reviews', *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France* (21 March 2018). https://doi.org/10.1145/nnnnnnn.nnnnnnn

15 Mortensen et al. (2017).

16 Wintle, Bonnie C., Mahlon C. Kenicutt II and William J. Sutherland. 'Scanning horizons in research, policy and practice', in *Conservation Research, Policy and Practice*, ed. William J. Sutherland et al. Cambridge University Press (in press).

17 McKinnon, Madeleine C. et al. 'Sustainability: Map the evidence', *Nature News, 528*(7581) (2015): 185. https://doi.org/10.1038/528185a; Sutherland and Wordley (2018).

18 Pamlin and Armstrong (2015).

19 Haddaway et al. (2015); Donnelly et al. (2018).

20 Donnelly et al. (2018).

21 Pamlin and Armstrong (2015).

22 Bostrom and Ćirković (2008).

23 Avin et al. (2018).

24 *Hands-On Machine Learning With Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media (2019).

25 O'Mara-Eves et al. (2015); Géron (2019).

26 e.g. Mortensen et al. (2017); Krivosheev, Casati, and Benatallah (2018).

27 Westgate et al. (2018).

28 e.g. Avin et al. (2018); Denkenberger, David C. and Joshua M. Pearce. 'Cost-effectiveness of interventions for alternate food in the United States to address agricultural catastrophes', *International Journal of Disaster Risk Reduction, 27* (1 March 2018): 278–89. https://doi.org/10.1016/j.ijdrr.2017.10.014

29 Wallace et al. (2013); O'Mara-Eves et al. (2015).

30 Rees (2003).

31 Pamlin and Armstrong (2015).

32 Baum, Seth and Anthony Barrett. 'Towards an integrated assessment of global catastrophic risk', *First International Colloquium on Catastrophic and Existential Risk: Proceedings* (2017): 53–80.

33 Avin et al. (2018).

34 Bostrom (2013).

35 Sutherland and Wordley (2018); Sutherland, William J. et al. 'Building a tool to overcome barriers in research-implementation spaces: The conservation evidence database', *Biological Conservation, 238* (2019). https://doi.org/10.1016/j.biocon.2019.108199

36 Donnelly et al. (2018).

10. The Mortality of States (MOROS) Dataset

Luke Kemp

Highlights:

- Having better data on states and their lifespans can help us understand both the phenomenon of collapse and the nature of entities that dominate global risk.
- This chapter documents the creation of a database of state lifespans, where a state is defined as "a set of centralised institutions that coercively extract resources from, and impose rules on, a territorially circumscribed population" and their lifespan is defined by "rough, critical dates in which significant changes to state form, function, and/or sovereignty occurred".
- The database was synthesised from a variety of primary data sources verified and expanded with a wider literature review.
- Significant interpretation was required to conceptualise states and their lifespans, but efforts were made to make this consistent and objective, while recognising that it is ultimately a qualitative overview of expert opinion.
- In future it is hoped to use expert elicitation and structured literature reviews to improve the database, alongside finding better ways to code for the continuity of states and adding details about the consequences and reasons for state termination.

https://doi.org/10.11647/OBP.0360.10

This chapter lays out a work-in-progress developmental methodology to measure the longevity of states. Arguing that the state, within the international, is an under-theorised object in Global Catastrophic Risk Studies, the chapter proposes the value of a standardised dataset for enhancing how we understand the role of states in GCR production and mitigation. Further reflections on dataset creation and modelling can be read in Chapters 9 and 20, while an alternative approach to studying state collapse can be found in Chapter 13.

1. Background

Our world is dominated by political states. Collapse is, at heart, the fall of a state and global risks are largely produced by a small number of powerful states and state-backed corporations.[1] Despite this, states are dramatically understudied in the realm of Global Catastrophic Risks.

Having better data on states and their ends can help fill this gap. Currently there are few resources which provide an overview of the lifespans of different states as well as the theories on why some were fragile and others resilient. Instead, there is a patchwork of different datasets, ranging from data collected by Rein Taagepera which summarise the lifespan of a selection of empires,[2] the Seshat database of historical polities,[3] and the Correlates of War Project.[4] Each of these is limited in some way. The data from Taagepera is decades old and only covers a few dozen empires (focused mainly on Eurasia); the Seshat database is not explicitly focused on state termination and crisis, while the Correlate of War Project only covers states after 1815, a small slice of history.

Unsurprisingly, we know little about the lifespan of states or societies. While indispensable background to thinking about global risk, there have only been two studies. In the largest piece to date, Arbesman analysed 42 empires (covering the period 3000 BCE to 600 CE) finding an ageless distribution: the risk of termination was constant, and an empire was just as likely to end at age 20 as age 200. Another focusing on 22 Chinese Dynasties (221 BCE – 1912 CE) found a power-law distribution leading the authors to suggest that it was organised by self-organised criticality. Both projects are limited. The first overlooks empires after 600 CE and does not cover all empires during its period of focus. The latter focuses just on Chinese empires during a small time slice.

In a short piece I wrote for *BBC Future* in 2019, I gathered a larger dataset, building off the work of Arbesman.[5] While bigger, it was still incomplete. This piece highlights a new, more comprehensive and systematic way of analysing past state terminations: the Mortality of States (MOROS) dataset.

2. Overview

The Mortality of States Index (MOROS) provides an overview of the lifespan of different states. It documents state formation and end dates for over 440 different states, covering roughly 5,000 years from 3100 BCE (Egyptian Dynasties I and II) to 2021. We define the state as a set of centralised institutions that coercively extract resources from, and impose rules on, a territorially circumscribed population. This is a necessarily broad definition. There was significant variety in how pre-modern states governed, as well as their level of administration, centralisation, and coercion. It is not an on/off switch: statehood exists on a spectrum.[6]

The idea of the state is not without detractors. There are critiques including the sheer diversity of states, and that the idea of a state is an inappropriate false projection of modern polities onto ancient cases.[7] These are not compelling arguments and are usually aimed at a strawman definition: Weber's outdated idea that the state is a monopoly on violence.[8] Social scientists have crafted sufficiently wide definitions (like the one used earlier) which can capture both Pharaonic Egypt and the modern US. Nonetheless, there is general consensus across political science and archaeology that, despite significant variety, states provide a real and useful political category. While there are difficulties in drawing precise boundaries it is a commonly used category in the social sciences.[9]

States are far more easily measured than more amorphous concepts such as society and civilisation. There are few accepted definitions of "civilisation" or "society", and determining a beginning and end date is even more pernicious. States represent a more easily defined, measured, and concrete historical unit.

The dates for both state beginnings and ends should be seen as rough, critical dates in which significant changes to state form, function, and/or sovereignty occurred. They are often indicative of processes that may have taken decades to unfold. For instance, 1177 is used as rough dating for the collapse of the Bronze Age state network, even though the process unfurled over decades.[10]

3. Methods

The entries have been gathered from a range of different materials. I worked with Oscar Rousham (see "Acknowledgements") to pull together the first dataset. The initial, primary sources were:

- Three different surveys of historical empires and large polities by Taagepera;[11]
- The Seshat Database;[12]
- The four-volume 2016 *Encyclopedia of Empire*[13] and;
- The Correlates of War Project.[14]

We used these primary sources to generate the first dataset. We compared entries to eliminate duplicates. Seshat encompasses not only states, but broader cultural periods that are distinguished by changes in material culture. Hence, entries from Seshat were only included when they represented a distinct and established state.

Each of the primary sources focuses on overlapping, similar units, although with differing definitions. Most sources lack a distinct measurement of statehood, and hence a guide to coding for state formation, continuity, and termination. The sole exception is the Correlates of War Project which uses both political recognition from great powers and a population size of at least half a million as proxies for state sovereignty. While more specific, the arbitrary population threshold is not appropriate for pre-modern states with often significantly lower populations, and recognition by neighbours would inappropriately exclude many ancient (especially "pristine") states.

We then drew on a wider literature search to both verify the majority of existing entries and to create an additional 22. Most of these were from speciality sources for Chinese dynasties[15] and Korean kingdoms[16] which were less covered by primary sources. We also consulted books focused on societal collapse, although these were either unsuitable or already included.[17]

Where we have found competing suggestions on state formations and termination dates, we input both the lowest and highest credible estimates.

During this construction phase we excluded 30 polities. Entries were excluded for one of two reasons: a) it was unclear whether the polity would qualify as a state, and/or b) the formation and termination dates were highly uncertain (spanning decades) and/or contested. These are Benin Edo, the Brunei Sultanate, Chavin, Da Viet, Elam, Indus Valley Civilisation, Hurrian Kingdoms of Urkesh, Jene Jano, Kanem Bornu, the Maya, Minonan Crete, the Moche, Mutapa, Ndebele, Ngoni, Papal States, Rapa Nui, Republic of Pisa, Shona, Srivijaya Empire, Teutonic Order, Tui-Tonga, Tukolor Order, Venetian Empire, Vishnukundina Dynasty, Wahabi Empire, Western Satrap, Xianbei, Yap Empire, and Zapotec. We expect that many of these could be clarified and included in a future version of MOROS.

We applied four criteria to assess statehood:

- The presence of a state apparatus that was formally (and even legally) capable of imposing rules.
- Institutionalised authority that could enact the functions of the state without relying on the charisma of the ruler.
- Continuous rule over a territory extending beyond a single city.
- The level of expert (dis)agreement.

The dates in MOROS do not represent any quantitative thresholds. Instead, they represent rough agreement by experts as to when a state can be said to have existed and ended based on interpretation of an array of sources and factors. It is a qualitative overview of common expert opinion on political periodisation. This poses problems. Different experts, and different fields can implicitly deploy varying interpretations of what signifies the end of a polity or lineage. This is difficult to detect since experts frequently do not explicitly define state formation and termination. Nonetheless, this approach remains a credible way of determining state formation and termination.

Note that the dates provided in MOROS say little about the exact nature of the state formation and end. An empire in the dataset could have undergone a full collapse of political, economic, and societal institutions, or just undergone a fundamental change in political form (such as the movement of Rome from Republic to Empire). It also covers a simpler change in ruling elites, such as dynastic shifts in China that were incurred by internal warlords or coups (which we have identified and marked within MOROS). We hope to use expert elicitation and systematic literature reviews in the future to provide deeper information on the exact details of each entry, including what the termination entailed, a stricter definition of state formation and termination, the purported causes for collapse/transformation, and the evidence underpinning different theories.

MOROS is a work in progress. Further work is needed to ensure the estimates, are robust, comparable, and provide appropriate depth in analysis. It is not entirely comprehensive of either all states throughout human history or for all types of polities. It excludes city-states, non-state polities, and more amorphous units such as “civilisations”. Nonetheless, it is — to the best of our knowledge — the largest dataset of state lifespans in existence.

4. Next Steps

MOROS is a provisional tool. There are many promising ways to expand and refine it. The first and most pressing is to simply find better ways to code for the continuity of states. The current dataset simply depicts the most accepted historical chronologies, although these do not have a common definition for state termination and formation. Coding the data with more strict definitions (such as a prolonged loss of sovereignty) would be a more robust and consistent approach.

A second path is to look beyond simple dates towards the consequences, and the reasons for termination. Separating genuine cases of collapse from simple elite replacement would make this a far more useful tool, as would detailing the proximate and ultimate reasons for termination. Pairing MOROS with expert elicitation and a literature review are two methods to include reasons and consequences into MOROS.

Acknowledgements

I would like to acknowledge funding from the Berkeley Existential Risk Initiative which allowed for Oscar Rousham to help in collecting the data.

Notes and References

[1] Kemp, Luke. 'Agents of doom: Who is creating the apocalypse and why', *BBC Future* (2021). https://www.bbc.com/future/article/20211014-agents-of-doom-who-is-hastening-the-apocalypse-and-why?ocid=twfut

[2] Taagepera, Rein. 'Size and duration of empires: Growth-decline curves, 600 BC to 600 AD', *Social Science History, 3*(3) (1979): 115–38. https://doi.org/10.1017/S014555320002294X; Taagepera, Rein. 'Expansion and contraction patterns of large polities: Context for Russia', *International Studies Quarterly, 41*(3) (September 1997): 475–504. https://doi.org/10.1111/0020-8833.00053; Taagepera, Rein. 'Size and duration of empires: Growth-decline curves, 3000 to 600 BC', *Social Science Research, 7*(2) (June 1978): 180–96. https://doi.org/10.1016/0049-089X(78)90010-8

[3] Turchin, Peter et al. 'Seshat: The global history databank', *Cliodynamics: The Journal of Quantitative History and Cultural Evolution, 6*(1) (4 July 2015). https://doi.org/10.21237/C7CLIO6127917; Seshat. *Seshat Code Book* (2021). http://seshatdatabank.info/wp-content/uploads/2021/04/Code-Book-4.20.2021.pdf

[4] COWP. *Correlates of War Project* (2021). https://correlatesofwar.org/

[5] Kemp, Luke. 'Are we on the road to civilisation collapse?', *BBC Future* (19 February 2019). https://www.bbc.com/future/article/20190218-are-we-on-the-road-to-civilisation-collapse

[6] Scott, James C. *Against the Grain: A Deep History of the Earliest States*. Yale University Press (2017); Scheidel, Walter. 'Studying the state', in *The Oxford Handbook of the State in the Ancient Near East and Mediterranean* (1st edition), ed. Peter Fibiger Bang and Walter Scheidel. Oxford University Press (2013): 5–58. https://doi.org/10.1093/oxfordhb/9780195188318.013.0002

[7] Graeber, David and David Wengrow. *The Dawn of Everything: A New History of Humanity*. Allen Lane (2021).

[8] Scheidel, Walter. 'Resetting history's dial? A critique of David Graeber and David Wengrow, "The Dawn of Everything: A New History of Humanity"', *Cliodynamics: The Journal of Quantitative History and Cultural Evolution* (28 April 2022). https://doi.org/10.21237/C7CLIO0057266

[9] Scheidel (2013).

[10] Cline, Eric. *1177: The Year Civilization Collapsed* (2nd edition). Princeton University Press (2021).

[11] Taagepera (1978); Taagepera (1979); Taagepera (1997).

[12] Turchin et al. (2015); Seshat (2021).

[13] Dalziel, Nigel and John M. MacKenzie (eds.). *The Encyclopedia of Empire*. John Wiley & Sons, Ltd (2016). https://doi.org/10.1002/9781118455074

14 COWP (2021).

15 Wilkinson, Endymion Porter. 'Chinese history: A new manual (5th edition)', *Harvard-Yenching Institute Monograph Series 100*. Harvard University Asia Center, for the Harvard-Yenching Institute (2018).

16 Lee, Soyoung, Denise Patry Leidy and Metropolitan Museum of Art (eds.). *Silla: Korea's Golden Kingdom*. The Metropolitan Museum of Art (2013).

17 Tainter, Joseph A. *The Collapse of Complex Societies*. Cambridge University Press (1990); Middleton, Guy. *Understanding Collapse: Ancient History and Modern Myths*. Cambridge University Press (2017).

11. Enabling the Participatory Exploration of Alternative Futures With ParEvo

Rick Davies, SJ Beard, Tom Hobson and Lara Mani

Highlights:

- ParEvo is an online method of developing alternative future scenarios using a participatory evolutionary process. This involves the reiteration of variation, selection, and reproduction of possibilities, i.e. an embodiment of the evolutionary algorithm.[1]
- The process is designed to be used by multiple people to produce a collective good — a set of storylines. In addition, the process generates data on the structure of participation — how people have collaborated to produce those storylines.
- For users, ParEvo can achieve two related purposes. The first is cognitive: to enable participants to creatively think about alternative futures and to prompt how they do that thinking (metacognition). The second is more behavioural: to prompt consideration of ways of responding to possible futures, in anticipation and/or in response, and to exploit and/or mitigate their consequences.

https://doi.org/10.11647/OBP.0360.11

- This chapter explored the origins and use of ParEvo and illustrates how it has been implemented and analysed. Illustrative references are made to two recent exercises carried out by CSER.
- Three types of evaluation challenges are highlighted, concerning the performance of individual participants, exercises, and the platform as a whole. Researchers are invited to explore the uses of the application and to address some of the challenges raised in this chapter.

Exploring possible alternative futures is an invaluable means for thinking about how risks and vulnerabilities might develop, and how they may be mitigated. At the same time, this type of collaborative, exploratory futuring exercise is also useful for illustrating that, while some path-dependencies should be attended to, catastrophic (or utopian) visions of the future are not ineluctable or inevitable — change always remains possible and pathways to a safer, more survivable future can always be taken. The opening up of these possibilities is explored, in a rather different way, in Chapter 21, while alternative means of exploring futures collaboratively are proposed in Chapters 8 and 15.

The Origins of ParEvo

The design of the ParEvo process had its origins in the lead author's 1998 PhD thesis on organisational learning within non-governmental aid agencies (NGOs). That conception of organisational learning was based on an evolutionary epistemology.[2] The same research led to the design of another method also using a social embodiment of the evolutionary algorithm known as Most Significant Change (MSC).[3] Now used widely for impact monitoring and evaluation in complex development projects,[4] MSC is a convergent and optimising process in contrast to ParEvo which is more divergent and satisficing.[5]

Uses to Date

As of mid-2022, 19 different ParEvo exercises have been completed, during and since the development of the web application.[6] Participants have included school students, volunteers recruited from evaluation communities of practice, crowdsourced paid adult university educated UK participants, staff from a UK development aid think tank, UN Volunteers, UN agency staff members, and internationally recognised experts in particular fields. Futures explored include post-Brexit Britain, climate change post-COP26, post-Trump USA, a five-year corporate strategy, uptake pathways for educational research, the global governance of biotechnology research risks, and the future of Existential Risk Studies. Alternative histories have also been explored, including agricultural development project implementation, UNV volunteer experiences, and gender policy implementation within a UN agency. Eight of the earlier exercises were initiated by the lead author; 11 of the more recent exercises were initiated by members of other interested organisations. Two of these were implemented by the Centre for the Study of Existential Risk (CSER), in Cambridge, UK, and have been used as illustrative examples. Four other exercises are now scheduled for 2022–2023.

How a ParEvo Exercise Works

3.1 The generation of storylines

Via an online interface at ParEvo.org, participants are presented with a seed text equivalent to the first paragraph of a novel. This text has been prepared by the exercise facilitator. Participants are then each given the opportunity, independently and anonymously in parallel, to extend that narrative by adding a following paragraph, describing what happened next. They are then allowed to view each of those alternative extensions, and then choose only one of those which they would most like to develop by adding another following paragraph, again independently and anonymously in parallel. In most exercises one immediate result will be that some initial versions of the story will be ignored, while others might be extended by more than one participant. As this process is reiterated what emerges is a branching tree structure of alternative storylines like that in Figure 1 below.

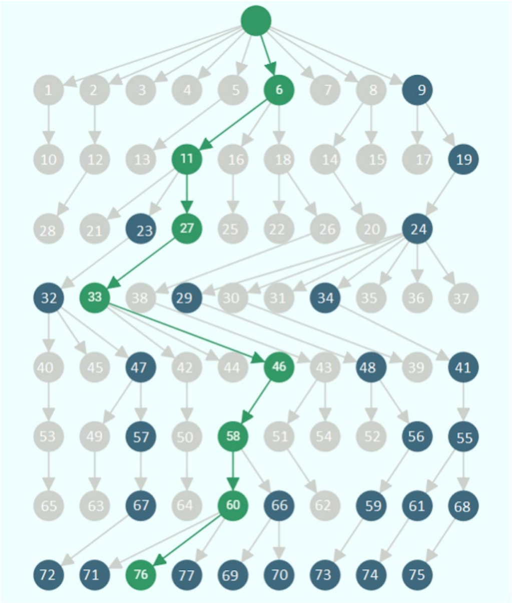
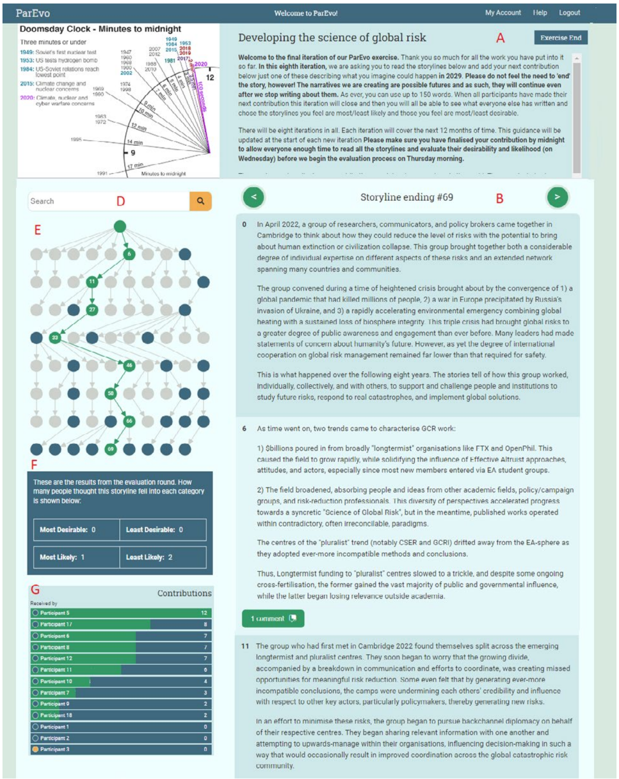


Fig. 1: Tree structure of alternative storylines generated by CSER Exercise 2.

In this CSER example there were up to 10 participants, who participated in eight iterations, visible as rows. The process begins at the top with a single seed paragraph and ends at the bottom, with nine surviving storylines. Each node represents a paragraph of text contributed by a participant, and the connecting lines show which new paragraph was added to which pre-existing paragraph. Grey nodes in the tree structure indicate paragraphs which were not continued and which in effect represent "extinct" storylines. Participants in subsequent iterations were not allowed to build on these. Dark green nodes represent paragraphs which others did build on and which became part of storylines which survived until the end of the exercise (bottom row). Bright green nodes represent one storyline which has been highlighted by a user of the ParEvo app. Doing so then brings up the full text of that storyline in a panel to the right of the tree structure, on the ParEvo user interface (seen in Figure 2 below).[7]



Developing the science of global risk

Welcome to the final iteration of our ParEvo exercise. Thank you so much for all the work you have put into it so far. **In this eighth iteration,** we are asking you to read the storylines below and add your next contribution below just one of these describing what you imagine could happen **in 2029. Please do not feel the need to 'end' the story, however! The narratives we are creating are possible futures and as such, they will continue even after we stop writing about them.** As ever, you can use up to 150 words. When all participants have made their next contribution this iteration will close and then you will all be able to see what everyone else has written and chose the storylines you feel are most/least likely and those you feel are most/least desirable.

There will be eight iterations in all. Each iteration will cover the next 12 months of time. This guidance will be updated at the start of each new iteration **Please make sure you have finalised your contribution by midnight to allow everyone enough time to read all the storylines and evaluate their desirability and likelihood (on Wednesday) before we begin the evaluation process on Thursday morning.**

Storyline ending #69

0 In April 2022, a group of researchers, communicators, and policy brokers came together in Cambridge to think about how they could reduce the level of risks with the potential to bring about human extinction or civilization collapse. This group brought together both a considerable degree of individual expertise on different aspects of these risks and an extended network spanning many countries and communities.

The group convened during a time of heightened crisis brought about by the convergence of 1) a global pandemic that had killed millions of people, 2) a war in Europe precipitated by Russia's invasion of Ukraine, and 3) a rapidly accelerating environmental emergency combining global heating with a sustained loss of biosphere integrity. This triple crisis had brought global risks to a greater degree of public awareness and engagement than ever before. Many leaders had made statements of concern about humanity's future. However, as yet the degree of international cooperation on global risk management remained far lower than that required for safety.

This is what happened over the following eight years. The stories tell of how this group worked, individually, collectively, and with others, to support and challenge people and institutions to study future risks, respond to real catastrophes, and implement global solutions.

6 As time went on, two trends came to characterise GCR work:

1) $billions poured in from broadly "longtermist" organisations like FTX and OpenPhil. This caused the field to grow rapidly, while solidifying the influence of Effective Altruist approaches, attitudes, and actors, especially since most new members entered via EA student groups.

2) The field broadened, absorbing people and ideas from other academic fields, policy/campaign groups, and risk-reduction professionals. This diversity of perspectives accelerated progress towards a syncretic "Science of Global Risk", but in the meantime, published works operated within contradictory, often irreconcilable, paradigms.

The centres of the "pluralist" trend (notably CSER and GCRI) drifted away from the EA-sphere as they adopted ever-more incompatible methods and conclusions.

Thus, Longtermist funding to "pluralist" centres slowed to a trickle, and despite some ongoing cross-fertilisation, the former gained the vast majority of public and governmental influence, while the latter began losing relevance outside academia.

1 comment

11 The group who had first met in Cambridge 2022 found themselves split across the emerging longtermist and pluralist centres. They soon began to worry that the growing divide, accompanied by a breakdown in communication and efforts to coordinate, was creating missed opportunities for meaningful risk reduction. Some even felt that by generating ever-more incompatible conclusions, the camps were undermining each others' credibility and influence with respect to other key actors, particularly policymakers, thereby generating new risks.

In an effort to minimise these risks, the group began to pursue backchannel diplomacy on behalf of their respective centres. They began sharing relevant information with one another and attempting to upwards-manage within their organisations, influencing decision-making in such a way that would occasionally result in improved coordination across the global catastrophic risk community.

These are the results from the evaluation round. How many people thought this storyline fell into each category is shown below:

Most Desirable: 0	Least Desirable: 0
Most Likely: 1	Least Likely: 2

Received by	
○ Participant 5	12
○ Participant 17	8
○ Participant 6	7
○ Participant 8	7
○ Participant 12	7
○ Participant 11	6
○ Participant 10	4
○ Participant 7	3
○ Participant 9	2
○ Participant 18	2
○ Participant 1	0
○ Participant 2	0
● Participant 3	0

Fig. 2: The ParEvo user interface (contributors' text obscured).

Figure 2 shows six other features of the user interface. These include guidance provided by the facilitator, updated with each new iteration, seen on the top right; the full text of a selected storyline, presented on the right; comments on two of the participants' contributions; a search facility, above the tree structure; an evaluation widget below the tree structure; and a "leader board" on the bottom left. The use of the search facility, comment facility and leader board are optional. The leader board was not used in the CSER exercises.

3.2 Evaluation of storylines

When the process of generating a set of storylines is finished, participants are then asked to evaluate the storylines that they have helped to generate. This can be done using the widget built into the app, shown in Figure 2, but can also be done using a more detailed online survey. The widget allows participants to make polar evaluation choices, i.e., which storyline they think is most or least likely, desirable, equitable, sustainable, etc. The available choices are set by the exercise facilitator. The online surveys used to date have used both open- and closed-ended questions about the contents and process (see example survey in Appendix 1). In addition, the facilitator can download from ParEvo.org 12 Excel formatted datasets containing automatically generated information about the contents of the storylines and the structure of people's participation in the exercise (listed in Appendix 2). Exercise facilitators have also organised post-exercise meetings of participants to solicit further views from the participants and to provide feedback on analyses by the facilitator.

3.3 Theory

There are three different types of theories about what happens during a ParEvo exercise. The first is the participants' own often tacit and informal theories about what might happen in the future, as evident in the contents of their contributions, and the resulting composite storylines.

The second is the exercise facilitator's expectations of what they want to see happen in the exercise they have designed and organised. As evident in the kind of futures they want to see explored, the kinds of people to be involved, the time span and granularity of the exercise, and the guidance they give to participants at the beginning of each new iteration. Exercise facilitators have a range of exercise design parameters they can vary in pursuit of these expectations (see Appendix 3).

The third is the expectations of the platform administrator (and designer of the ParEvo app). Each consecutive exercise has been, in effect, an opportunistic experiment, usually involving some new variations in the design parameters, primarily under the control of the exercise facilitator. Some of these variations are persisting across multiple exercises and others not. In addition to the evolutionary epistemology at the base of the design, there is an associated ongoing interest in the role of diversity. This perspective has been informed by writings on diversity from a complexity perspective,[8] measures of diversity used in ecology and sociological uses of those ecological ideas,[9] and network analysis as a way of visualising and measuring diversity measures.[10] Measurement has been relevant when thinking about diversity as an independent variable affecting the creativity of the process, but also as a dependent variable that is descriptive of the range of possible futures that have been developed. Underlying the design of ParEvo, when used to look forward, is the assumption that the generation and analysis of a diversity of storylines will enable participants to be better prepared for the future, which is only likely to be partially knowable at best. In this context, the intention is not to predict the future, but to be able to be more adaptive and responsive to the futures that might take place.

This approach is consistent with a substantial body of evidence on the importance of diversity to the more general task of effective problem-solving, as referred to recently in Campbell et al.[11] However, in this instance the ambition has its origins in the author's work as an evaluation consultant, assessing the performance of international development aid programmes, and the theories of change embedded in those programmes. These are characteristically optimistic, focusing on expected and desired futures, but are challenged by unexpected events and diverse implementation contexts.[12]

A small number of other types of online platforms have been developed by futurists for collaborative exploration of alternative futures, and subject to review.[13] ParEvo differs from these in three respects. Firstly, there is a generative theory informing the process design — the social embodiment of the evolutionary algorithm. Secondly the process is more divergent and satisficing rather than convergent and optimising, as is the case with online Delphi exercises — a widely used method. Thirdly, with ParEvo the construction of narratives precedes and informs a detailed analysis, rather than following and being informed by a technical analysis of other available data. In this respect, it is more ethnographic in orientation, taking participants' views as the primary resource material. Notwithstanding these differences, Raford's review of these platforms has provided a useful set of performance criteria relevant to the ongoing assessment of ParEvo and its further development (Appendix 4), some of which are discussed below.

Challenges

There are three broad challenges facing ParEvo facilitators and the administrator, including: (a) choosing the right design settings, (b) analysing the completed exercise, both the storyline contents and participation data, and (c) evaluating outcomes and impacts. The latter has been identified as an area of weakness in the field of futures research and practice[14] and will be explored here.[15] Evaluation can be done at three levels of aggregation, corresponding to the different types of theories of change introduced above: those of individual participants, individual exercises and the platform as a whole.

4.1 Participants

Until now participants have engaged in ParEvo exercises without receiving any intentional and explicit feedback on the nature of their individual performances. Nevertheless, evaluation surveys of participants in the last two CSER exercises indicate that many participants have enjoyed taking part.[16] Dropout rates during both exercises were small, with 93% of all 88 expected contributions being made in Exercise 1, and 96% of all 80 expected contributions being made in Exercise 2.

At best, participants can see what happens to their own contributions in subsequent iterations, i.e., whether one or more of the other participants choose to build on those contributions, and the way they do so — taking the storyline in the same direction or changing it radically. This behaviour seems to have different significance to different participants. The post-exercise surveys found varying opinions between the two exercises and within each exercise, with most but not all participants in the second exercise giving more importance to building on other participants' contributions, and vice versa.

Expectations and motivations might be expected to be different if there was more explicit feedback on participants' contribution behaviour. Facilitators now have the option of making a "leader board" visible on the user interface, which shows for each participant how many other participants have built on their contributions and those of others. This could have the effect of more directly motivating participants to seek these types of responses. Its consequences have not yet been tested, but one possibility is that it may lead to more convergent content. That may be desirable in some situations, as discussed in the next section on exercise level performance. Another leader board is under development where the performance measure is the proportion of all the contributions to the surviving storylines that were made by each participant. This will provide a more summative view of each person's contributions towards a more collective end. But again, its consequences have yet to be tested. Both possibilities can be seen as a form of gamification,[17] an approach already recognised as relevant to enabling collective intelligence.[18]

Other forms of more individualised and nuanced feedback are already available using the comment and tagging facilities, used after contributions have been made in each iteration. In exercises to date, including the recent CSER exercises, the comment facility has been only used in a very non-directive way by the facilitators, raising questions rather than proposing directions or signalling approval or disapproval. The option also exists for participants to (anonymously) comment on each other's contributions, and for this to affect the overall structure and direction of the storylines. Analysis of any evaluative content of this kind, and its influence, will be more challenging.

4.2 Exercises

The expected post-exercise impacts of a ParEvo exercise vary from exercise to exercise, depending on the individual facilitator's objectives. To date, these have included:

- Influencing the content of a strategic plan (one exercise completed, one planned)
- Informing the publication of papers in an academic journal (two exercises)
- Informing the content of an evaluation (four exercises)

- Leading to the revision of risk management protocol (one exercise in process)
- Changing plans for ensuring research uptake (one exercise in process)

There are constraints on the extent to which the specifics of these impacts, and the associated causal mechanisms, can be identified by the platform administrator. Facilitators are not obliged to share post-exercise survey data, or other information about the subsequent effects of their exercise.

However, it is possible to identify more immediate differences in exercise outcomes, as distinct from post-exercise impacts, using measures that can be applied to almost all exercises, regardless of their specific objectives. As mentioned in the section on theory above, the generation of a diversity of alternative futures has been a default expected outcome for almost all ParEvo exercises to date. Drawing from the field of ecology, Stirling differentiated three facets of diversity, each of which are measurable:

- variety, also known as richness, which is the number of different kinds, e.g. species;
- balance, also known as evenness, being the relative numbers of each kind; and
- disparity, the degree of difference between kinds, e.g. between people and chimpanzees versus people and bacteria.[19]

In the analysis of a number of ParEvo exercises, these aspects of diversity have been measured in three ways.

4.2.1 Tree structures

The first method looks at the network structure of the storylines, in terms of disparity. Some storylines are more similar than others, in that they have many contributions in common, only diverging in the last iteration. The content of others are less so because they diverged in the very first iteration. Comparing the structure of storylines in two CSER exercises (Figure 3 below), four of the original storylines survived to the last iteration in Exercise 1, but only two did so in the second. This aspect of diversity can be measured more specifically by counting the links connecting the surviving storylines, a simple network analysis measure of distance. There were 53 in Exercise 1, versus only 30 in Exercise 2. The significance of disparity is discussed further below.

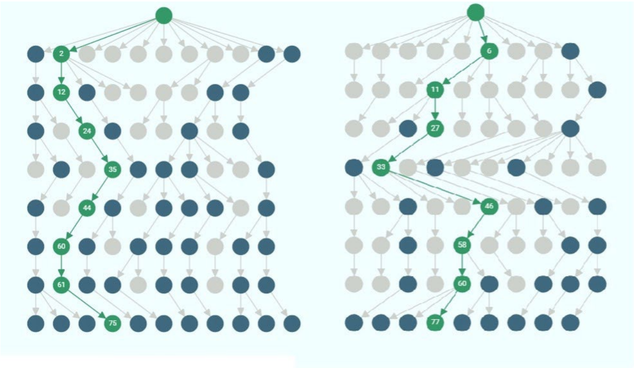


Fig. 3: Tree structures for CSER Exercises 1 and 2.

4.2.2 Combinations of sets of ideas

The second approach looks at the kinds of combinations of ideas that occurred, as a percentage of all possible combinations. Each participant can be seen as a set of ideas. When one participant's contribution builds on the contribution of another, this represents one of those kinds of combinations taking place. The total number of possible types of combinations can be seen in a participant *x* participant matrix, but is limited further if there are not enough iterations for all to occur, and if any participants drop out of any iterations. The percentage of those possible combinations actually occurring in Exercise 1 was 61%, whereas in Exercise 2 it was 70%. This measure, known as network density, is a crude measure of "variety".[20] This measure is of interest because the recombination of ideas is considered an important source of creativity both in biological evolution and human culture.[21]

4.2.3 Storyline evaluations

A third approach looked at diversity in the evaluation judgements of participants. At the end of both CSER exercises, after eight iterations had been completed, participants were asked to identify which specific surviving storylines they saw as describing the "most likely", "least likely", "most desirable", and "least desirable" futures — as seen from their own perspective. Their responses were then used to create a scatter plot within which there were four different quadrants of possibilities, as shown in Figure 4 below.

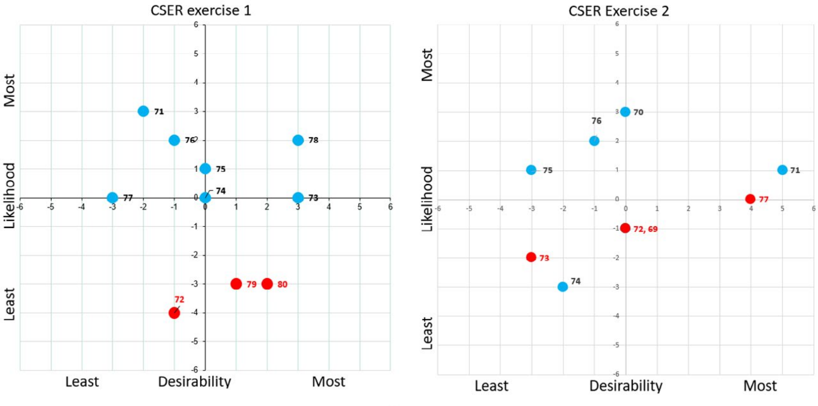


Fig. 4: Distribution of evaluations of the surviving storylines in CSER Exercises 1 and 2.

The values on each axis represent the number of participants who made judgements of that kind.[22] The assessment of storylines located towards the edges of the scatterplot had the support of more participants than those towards the centre. The red storylines were those where participants had conflicting judgements, where both polar extremes were applied to the same storyline, e.g. being most and least desirable. In Exercise 1 two of the three contradictory judgements were about desirability. In Exercise 2 three of the four contradictory judgements were about likelihood.

Diversity in this context can be measured in two different ways. Minimal diversity of judgement would be visible in the presence of only two storylines in the scatter plot, when all participants agreed that one storyline was most desirable and least likely, and the other was least desirable and most likely.[23] Maximum diversity would be visible where all the surviving storylines appeared on the scatterplot. In both CSER exercises there was maximum diversity on this variety measure of diversity. Within all the possible combinations of judgments, the most disparate would be where participants expressed contradictory judgements about the desirability, or the likelihood, of a storyline. As noted above, these kinds of judgements were seen in both exercises, slightly more so with Exercise 2.

4.2.4 Implications for analysis

The diversity measurement options just discussed can be seen as mediating variables possibly affecting the post-exercise impacts exercise facilitators are aiming for, such as those listed above, or as dependent variables, of interest as more proximate outcomes. In both cases the exercise settings can be seen as the independent variables. The relationship between these types of variables remains to be explored. Findings could then inform how future facilitators can optimise the design of their exercises.

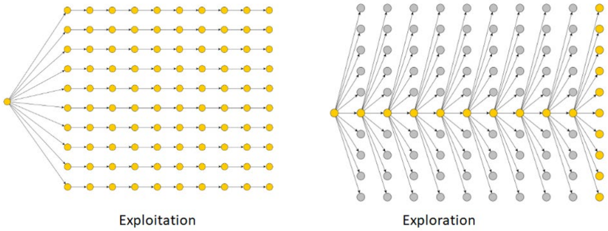


Fig. 5: Extreme examples of exploitation versus exploration search strategies in ParEvo exercises.

One dimension of optimisation is captured by the distinction made by organisational learning theorist James March, between "exploration", of the new, and "exploitation", of what is already known.[24] The question of which strategy is most appropriate in what conditions has been the subject of ongoing research since then.[25] The same distinction can be used to differentiate search processes used in different ParEvo exercises. The two extreme manifestations of these are shown in Figure 5. Exploration involves development of many disparate storylines, where exploitation involves the more detailed development of versions of a single preferred storyline.

Viewed from this perspective, and given the earlier analysis of disparity, Exercise 2 participants seemed to be pursuing a slightly more exploratory futures search than those in Exercise 1, although since participants were unaware of one another's choices, and thus could not know for sure whether they were building on the same contributions as their peers or not, this is hard to tell.

In the future, it is possible that some facilitators of ParEvo exercises may want to give more emphasis to exploitation, and for their participants to converge on a more specific view of the future. In the discussion above about the measurement of individual performance, two ways of measuring individuals' contributions were introduced, which if given as feedback via a leader board, could encourage such behaviour. That possibility needs to be tested.

4.5 Platform

At the platform level, objectives relate to sets of exercises. They have included:

- The development of new features in the ParEvo app. This was especially the case with six earliest exercises facilitated by the administrator, and has continued as a secondary objective thereafter. Visible improvements have included the development of a comment facility, a tagging facility, a leader board, and more flexible evaluation options.
- Further exercises by first-time facilitators. This has been the case with four facilitators, leading to six additional exercises to date.
- Requests by other organisations wanting to use ParEvo for the first time. Five new organisations were registered as users in 2022, with exercises planned in the next 12 months.[26]
- Accumulation of data from multiple exercises that is sufficient to enable analyses of exercise parameters and how they affect exercise outcomes. This process is underway. A supporting website, at https://mscinnovations.wordpress.com/, is accumulating information on the usefulness of different forms of analysis of this data.

- Publication of papers about ParEvo in academic journals and books. Six are in process to date.
- Recovery of investment costs, through payments for technical support provided to organisations using ParEvo. This has been underway since early 2022, although *pro bono* technical support for first-time users remains the norm, as does the free use of the app itself.

Objectives for the platform have changed over time. With the earliest exercises the main objective was to ensure that the web application functioned as expected. Then more attention was given to the development of evaluation options, within the app itself, and using third-party survey platforms. In the last 12 months, more emphasis has been given to ensuring sufficient post-exercise facilitated discussion amongst the participants, to work through the implications of the exercise, and its evaluation. This emphasis needs to be continued. More encouragement is also being given to exercise facilitators to articulate their objectives for their exercises before they start. Both facilitators did so in the two recent CSER exercises.

An Invitation

This chapter has provided a quick overview of ParEvo.org, a web-assisted process enabling the participatory exploration of alternative futures. It is hoped that researchers in Existential Risk Studies and elsewhere will see this as a potentially useful tool to explore how groups of people can collaboratively construct a diverse set of storylines around a topic of shared interest. Each exercise facilitator has considerable freedom in how they configure their own exercise. Lessons from previous exercises are available both from the ParEvo website and from the ParEvo administrator. Each exercise generates a significant body of qualitative and qualitative data, about both the storyline contents and participants' behaviour. A range of options already exist for the analysis of that data. Several important challenges relating to assessment of performance at different levels of aggregation have also been identified and could be addressed.

The following appendicies are available on-line 1) an example on-line survey, 2) a summary of downloadable datasets, 3) the adjustable parameters of a ParEvo exercise, and 4) a set of Platform Assessment Criteria.

https://doi.org/10.11647/OBP.0360#resources

Notes and References

[1] Campbell, D. T. 'Blind variation and selective retentions in creative thought as in other knowledge processes', *Psychological Review, 67*(6) (1960): 380–400. https://doi.org/10.1037/h0040373; Dennett, D. C. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. Penguin UK (1996).

[2] Campbell (1960); Dennett (1996); Bateson, G. *Mind and Nature: A Necessary Unity*. E. P. Dutton (1979).

[3] Davies, R. *Order and Diversity: Representing and Assisting Organisational Learning in Non-Government Aid Organisations* [PhD thesis]. University of Wales (1998). http://mande.co.uk/blog/wp-content/uploads/2013/05/thesis.htm; Davies, R. and J. Dart. *The 'Most Significant Change' (MSC) Technique: A Guide to Its Use* (2005). https://www.researchgate.net/publication/275409002

[4] Indirect evidence being the more than 36,000 reads of the 2005 MSC guide, recorded on https://www.researchgate.net/.

[5] Simon, H. A. 'Rational choice and the structure of the environment', *Psychological Review, 63* (1956): 129–38. https://doi.org/10.1037/h0042769

[6] For a full listing see https://mscinnovations.wordpress.com/list-of-parevo-exercises/.

[7] The participants' contributions have been blurred, because the publication of the text content of any ParEvo exercise is limited by the terms of the privacy policy, seen here https://parevo.org/privacy/.

[8] Page, S. E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*. Princeton University Press (2008); Page, S. E., E. Lewis, N. Cantor and K. Phillips. *The Diversity Bonus: How Great Teams Pay off in the Knowledge Economy*. Princeton University Press (2017).

[9] Stirling, A. 'A general framework for analysing diversity in science, technology and society', *Journal of the Royal Society Interface, 4*(15) (2007): 707–19. https://doi.org/10.1098/rsif.2007.0213

[10] Borgatti, S. P., M. G. Everett and J. C. Johnson. *Analyzing Social Networks* (2nd edition). Sage Publications (2018).

11 Campbell, C. M., E. J. Izquierdo and R. L. Goldstone. 'Partial copying and the role of diversity in social learning performance', *Collective Intelligence, 1*(1) (2022). https://doi.org/10.1177/26339137221081849

12 Davies, R. 'Representing theories of change: Technical challenges with evaluation consequences', *Journal of Development Effectiveness* (2018): 1–24. https://doi.org/10.1080/19439342.2018.1526202

13 Raford, N. 'Online foresight platforms: Evidence for their impact on scenario planning & strategic foresight', *Technological Forecasting and Social Change, 97* (2015): 65– 76. Hew, A., R. K. Perrons, S. Washington, L. Page and Z. Zheng. 'Thinking together about the future when you are not together: The effectiveness of using developed scenarios among geographically distributed groups', *Technological Forecasting and Social Change, 133* (2018): 206–19. https://doi.org/10.1016/j.techfore.2018.04.005

14 Gardner, A. L. and P. Bishop. 'Expanding foresight evaluation capacity', *World Futures Review, 11*(4) (2019): 287–91. https://doi.org/10.1177/1946756719866271

15 Other forthcoming papers by Beard, Mani and Hobson will explore the first two.

16 In both exercises, the average rating on an enjoyment scale (from 'Not at All' to 'A Lot') was above the midpoint.

17 Defined as the use of game features in non-game settings.

18 Cincilla, G., S. Masoni and J. Blobel. 'Individual and collective human intelligence in drug design: Evaluating the search strategy', *Journal of Cheminformatics, 13*(1) (2021): 80. https://doi.org/10.1186/s13321-021-00556-6; Riar, M., B. Morschheuser, R. Zarnekow and J. Hamari. 'Gamification of cooperation: A framework, literature review and future research agenda', *International Journal of Information Management, 67* (2022): 102549. https://doi.org/10.1016/j.ijinfomgt.2022.102549

19 Stirling, A. 'On the economics and analysis of diversity', *Science Policy Research Unit (SPRU), Electronic Working Papers Series, Paper, 28* (1998): 1–156.

20 Balance can also be calculated by examining the number of instances of each type of combination.

21 Xiao, T., M. Makhija and S. Karim. 'A knowledge recombination perspective of innovation: Review and new research directions', *Journal of Management, 48*(6) (2022): 1724–77. https://doi.org/10.1177/01492063211055982

22 Here, their scale position is a net value. The number of participants selecting a storyline as least likely (/least desirable) was subtracted from the number who selected the same storylines as likely (/desirable).

23 Or alternatively, most likely and most desirable, and least likely and least desirable.

24 March, J. G. 'Exploration and exploitation in organizational learning', *Organization Science, 2*(1) (1991): 71–87.

25 Wilden, R., J. Hohberger, T. M. Devinney and D. Lavie. 'Revisiting James March (1991): Whither exploration and exploitation?', *Strategic Organization, 16*(3) (2018): 352–69. https://doi.org/10.1177/1476127018765031

26 See https://mscinnovations.wordpress.com/list-of-parevo-exercises/.

III. RISK DRIVERS AND IMPACTS

The chapters in this section showcase a small number of the plentiful outputs that researchers at the Centre for the Study of Existential Risk have produced on the individual causes of extreme global risk: the natural phenomena, systemic shifts, and novel technologies that carry the potential to do massive harm at the global scale.

It is common practice in Existential Risk Studies to describe these as discrete existential risks or Global Catastrophic Risks, with categories that align rather neatly with the imagined boundaries of an issue: for example, AI, biotech, supervolcanoes, and climate change. However, as described in the introduction to this book, researchers at CSER and elsewhere are increasingly taking a different approach. There are many kinds of global and existential catastrophes and there are also many events, or chains of events, with the potential to bring these about. However, there is no simple one-to-one matching between these things, nor should their likeness in potential scale lead us to think that they require similar responses or even the same method of study or types of expertise. For instance, one kind of global catastrophe involves darkening of the earth's atmosphere, leading to reduced sunlight, crop failure, and global food insecurity. However, this is a catastrophe that could occur due to a variety of causes, including volcanic eruptions, asteroid impacts, nuclear war, or even geoengineering. Furthermore, it is a catastrophe that we could respond to in a variety of ways, by seeking to avoid such hazards from occurring (the most obvious response to the threat of nuclear war) or by planning for how we might still be able to feed everyone (or at least some people) even under such cataclysmic conditions (perhaps a more reasonable response to the threat from volcanic supereruptions that are far harder to prevent). Furthermore, this one kind of catastrophe is not the only thing that we need to worry about from each of these hazards.

For instance, Chapter 12, *Global Catastrophic Risk From Lower Magnitude Volcanic Eruptions* by Mani et al. argues that focusing only on the potential for volcanic eruptions to cause a volcanic winter means that the Existential Risk Studies community has equated Global Catastrophic Volcanic Risk with the risk from only the most explosive volcanic eruptions. However, volcanic eruptions can also have other global scale impacts, such as disruption to critical infrastructure, and these might easily be triggered by a far less explosive volcanic eruption were it to occur in the wrong place at the wrong time. Given this complexity, there is a real risk that in talking about specific global catastrophic risks we will overlook important catastrophe scenarios and the forces that could cause them.

A similar point is made in Chapter 14, *Existential Change* by Kemp and Beard, which argues that it is problematic to try and answer the question of whether something (in this case climate change) is or is not a Global Catastrophic or existential risk. There are many reasons we should not do this. For one thing, risk is a probabilistic concept, meaning that to provide a yes/no answer is inherently misleading. For another thing, risk drivers do not only operate independently but interact in both positive and negative ways. Thirdly, the very notions of existential risk and Global Catastrophic Risk are fuzzy and poorly defined, and often used in ways that elide their common-sense meanings. And finally, risks are not only things that happen to us but are also things that we actively create in choosing how we respond to events. Climate risk therefore involves the risk that we will respond to climate change in ways that are harmful instead of (or as well as) helpful, potentially making it much more dangerous. For all of these reasons then, the question we need to ask is not whether something is an existential or Global Catastrophic Risk, but what contribution it might make to the overall level of risk in one or more possible future scenarios. The chapters in this section all seek to answer this question in different ways, but are united in seeking to do so in a thoughtful and robust manner.

As already mentioned, Chapter 12 focuses on the contribution of volcanic eruptions. Drawing upon the global systems and vulnerability and exposure-based approaches to Existential Risk Studies described in Section 1, the authors argue that there has, to date, been too strong a focus on the possibility of direct volcanic hazards, in particular volcanic winters caused by the large amounts of material ejected into the atmosphere by the most explosive volcanic eruptions. They point out that historically it has not always been the largest volcanic eruptions that are the most damaging or costly, but rather those that are most disruptive to humans. To understand which volcanic eruptions might be maximally disruptive, the authors combine data about the location of active volcanoes with critical infrastructure, including cities, manufacturing centres, and key air and sea corridors through which people and goods most routinely travel. In this way they identify a number of “pinch points”, within which critical infrastructure and active volcanoes are closely co-located. Even a smaller volcanic eruption in one of these areas could have the potential to do significant harm at the global scale via cascading economic, social, and political impacts.

Chapter 13, *Reframing the Threat of Global Warming* by Richards et al., turns to the contribution of climate change. Once again, drawing on a global systems approach, the chapter seeks to understand Global Catastrophic Climate Risk, not merely in terms of the direct hazard from climate change but in terms of developing more complete catastrophe scenarios that take account of the cascading effects that climate change might have. The chapter focuses in particular on a cascade pathway connecting climate change to food insecurity and societal collapse. The authors conduct a systematic (manual) literature review of this topic and code studies based on the kinds of causal connections they investigate and the methods they use to do this. By looking at the results of studies they then construct an empirical causal loop diagram that presents the connections that are supported by evidence, their estimated strength and direction, and the quality and kind of research that has so far been used to study them. This helps both to understand the nature of this risk more fully and to plan further work to improve our, currently highly limited, understanding of this risk driver. A key finding of this study is that the global impacts of climate change require evidence and modelling at sub-global degrees of granularity, in order to account for more localised factors such as the distribution of populations and natural resources, the role of specific institutions and their potential failure, and the interaction between states and other groups through processes such as trade, migration, and conflict. If we only consider climate change scenarios at the global level, we will miss important considerations such as these, and are thus very likely to misunderstand the real contribution of climate change to global risk.

Chapter 14, *Existential Change: Lessons From Climate Change for Existential Risk*, provides a short addendum that draws some key lessons from other work at CSER on climate change. The chapter develops lessons for the wider Existential Risk Studies community from climate risk and how it has been studied in the past. Foremost among these is the importance of attending to questions of how different hazards and vulnerabilities might contribute to Global Catastrophic and existential risk, rather than focusing more exhaustively on questions of whether or not particular consequences of climate change do or don't constitute Global Catastrophic or existential risks. The chapter also offers some reflections on why climate change has been, to date, neglected within Existential Risk Studies (and why the most severe climate scenarios are also neglected within climate change research). Finally, the chapter suggests the importance of assessing response risks, and the potential co-benefits of risk mitigation.

Chapter 15, *A Fate Worse than Warming* by Tang and Kemp, considers the risk of Stratospheric Aerosol Injection (SAI). This is a technology that uses sulphate aerosols to alter the albedo of the upper atmosphere, reflecting more solar radiation out into space and thus reducing the amount of the sun's energy available to warm the surface of our planet. In theory, this could allow us to reverse some of the impacts of climate change on surface temperature; however, in doing so it will also have other impacts on the Earth's atmosphere and climate. The chapter seeks to perform a preliminary analysis of SAI as both a contributor and potential mitigator of Global Catastrophic Risk. A key element of the chapter is providing an analytical framework for different ways in which SAI might interact with Global Catastrophic Risk. These include its direct impacts, for instance its climatic and environmental effects, its interaction with other risk drivers such as nuclear war, the systemic risk it poses by stressing and/or changing both natural and social systems, and the latent risk posed by its potential termination and the negative impacts this might have.

Chapter 16, *Bioengineering Horizon Scan 2020* by Kemp et al., looks at biotechnology and provides a horizon scan of emerging issues in bioengineering. While not solely risk orientated, the chapter uses horizon-scanning techniques, as described in Chapter 6, to study possible future developments in the field. These cover a range of time frames, from issues predicted to emerge in the next five years, some of which (such as the rapid advances of automation in functional protein synthesis) have subsequently emerged as important issues in the field, to more distant issues that may still be a decade or more away but that should still be considered carefully and planned for. The study involves participants from six continents and covers both technical and governance specialists. Unsurprisingly it therefore highlights developments across the technology (from agricultural gene drives to neuronal probes), its governance (from the regulation of genetic databases to the governance of cognitive enhancement) and wider social and ethical trends (such as the increasing role of philanthropy in shaping research agendas or the potential misuses of neurochemistry) as equally important. Participants also identify seven underlying themes as driving change across bioengineering: political economy and funding; ethical and regulatory frameworks; climate change; transitioning from lab to field; inequalities; technological convergence; and misuse of technology, considering how these could be leveraged in ways that could make direct bioengineering research in more or less positive ways. Finally, the study highlights potentially fruitful further avenues for research, including focusing on specific areas of bioengineering, such as catastrophic risks; incorporating decision-support tools such as fault-trees; examining bioengineering in tandem with overlapping areas such as artificial intelligence; and producing policy-focused scans involving greater engagement with regulators.

Chapter 17, *Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI*, by Cremer and Whittlestone, looks at potential future developments in Artificial Intelligence. The chapter seeks to develop a methodology for identifying 'artificial canaries': that is, future developments in AI that could signal that we are approaching points at which its transformative potential may rapidly increase. The chapter chooses to focus on the transformative potential of AI, rather than more traditional metrics such as the level or generality of its intelligence, in recognition of the fact that AI is deeply enmeshed within socio-technological systems, and its contribution to Global Catastrophic and existential risk is likely to be mediated through this. Thus, a technology that fundamentally alters aspects of our economic or political landscape may contribute significantly to risk, even if it is not technically very different from its predecessors. The authors propose using a variety of participatory research methods, including workshops and structured expert elicitation exercises, to gather information about what future AI transformations might look like and how changes in the technology and its place in society could precipitate these. The authors further note that these methods should serve to promote more interdisciplinary research on these complex, multifaceted problems. Their methodology combines these participatory methods with collaborative causal graphs (similar to the empirical causal loops discussed in Chapter 13) to identify key milestones in the future development of AI and the relationships between them. They illustrate this approach with two examples, identifying artificial canaries for the use of AI in voter manipulation and developments towards High-Level Machine Intelligence. Their goals for this methodology are twofold. Firstly, identifying early warning signs of transformative applications of AI can support efficient monitoring and timely regulation of progress in its development. Secondly, those impacted by AI must have a say in how it is governed, and early warnings can give the public time and focus to influence emerging technologies using democratic, participatory processes.

All of the chapters contained in this section do much more than simply describing threats to humanity. They build on and utilise both the conceptual and philosophical insights from Section 1 and the methodological innovations from Section 2 to provide new ways of obtaining, synthesising, analysing, and presenting evidence about different drivers of risk. However, they also do so in a way designed to draw in decision-makers and help them to understand their own role in relation to extreme global risk. Whether it is appreciating how the geographical location of infrastructure or national level policies about institutional design, trade, migration, and defence can directly impact the global risk posed by volcanoes and climate change, the potential for response risks and complex technologies like SAI to cause well intentioned policies to do more harm than good, or the complex interactions between regulation, ethics, and technological development in bioengineering and AI, these chapters show the many ways in which policy-makers at many levels are influencing the most severe risk facing all of us. This research thus aims to be responsible both in seeking to improve the field as a whole, by moving beyond mere speculation to rigorous science, and also in positioning its findings in ways that support better policy making to try and reduce these risks.

In the final section of this book, we will turn to considering questions about policy engagement more deeply and shift our focus from understanding existential and Global Catastrophic Risk to understanding the best options available to us for promoting global safety, and existential hope.

12. Global Catastrophic Risk From Low Magnitude Volcanic Eruptions

Lara Mani, Asaf Tzachor and Paul Cole

Highlights:

- In terms of global-scale catastrophe, most research on volcanic risk has thus far focused on the impact of only the largest volcanic eruptions.
- However, this focuses only on the catastrophic potential of volcanic hazards and ignores issues of our exposure and vulnerability to them.
- A more systematic approach to Global Catastrophic Volcanic Risk highlights how globalisation has supported the clustering of critical infrastructure systems, sometimes in proximity to volcanic centres.
- These include areas around Taiwan, the Chinese-North-Korean border, the Luzon Strait, the Strait of Malacca, the Mediterranean sea, the North Atlantic and the Pacific Northwest.
- In this emerging risk landscape, even lower magnitude volcanic eruptions in these areas might have cascading, catastrophic effects and risk assessments ought to be considered in this light.

https://doi.org/10.11647/OBP.0360.12

This article was originally published in *Nature*, and highlights how a more systemic approach to thinking about extreme global risk, grounded in the existing sciences of volcanology, geography and sustainability, can reveal new risks and approaches to risk management. This chapter draws on the systemic approach to Global Catastrophic Risk described in Chapter 3.

Within the volcanic risk literature, the typical focus of attention for global-scale catastrophes has been on large-scale eruptions with a Volcanic Explosivity Index (VEI) of 7 to 8,[1] which remain relatively rare.[2] The relationship between volcanic eruptions of this scale and Global Catastrophic Risks (GCRs) — events that might inflict damage to human welfare on a global scale — provided rationality for this tendency.[3] We define this correlation as a "VEI-GCR symmetry", whereby as the magnitude of an eruption increases, so too does the probability of a GCR event. The eruption of Tambora in 1815 (VEI 7) is an example of the mechanism that governs the VEI-GCR symmetry, in which a large release of sulphur into the stratosphere brought about periodic global cooling, widespread frosts in the northern hemisphere, and crop failures across Europe.[4] This VEI-GCR symmetry has historically defined society's relationship with volcanoes. Indeed, we have often failed to consider lower-magnitude VEI eruptions as constituting GCRs.

Here, we argue that this symmetry has become imbalanced towards "VEI-GCR asymmetry", driven by a clustering of our global critical systems and infrastructures in proximity to active volcanic regions. Critical systems and infrastructures, such as shipping passages, submarine cables and aerial transportation routes, are essential to sustain our societies and to ensure their continued development.[5] We observe that many of these critical infrastructures and networks converge in regions where they could be exposed to moderate-scale volcanic eruptions (VEI 3–6). These regions of intersection, or *pinch points*, present localities where we have prioritised efficiency over resilience, and manufactured a new GCR landscape, presenting a new scenario for global risk propagation.

1. A Manufactured Global Catastrophic Risk Landscape

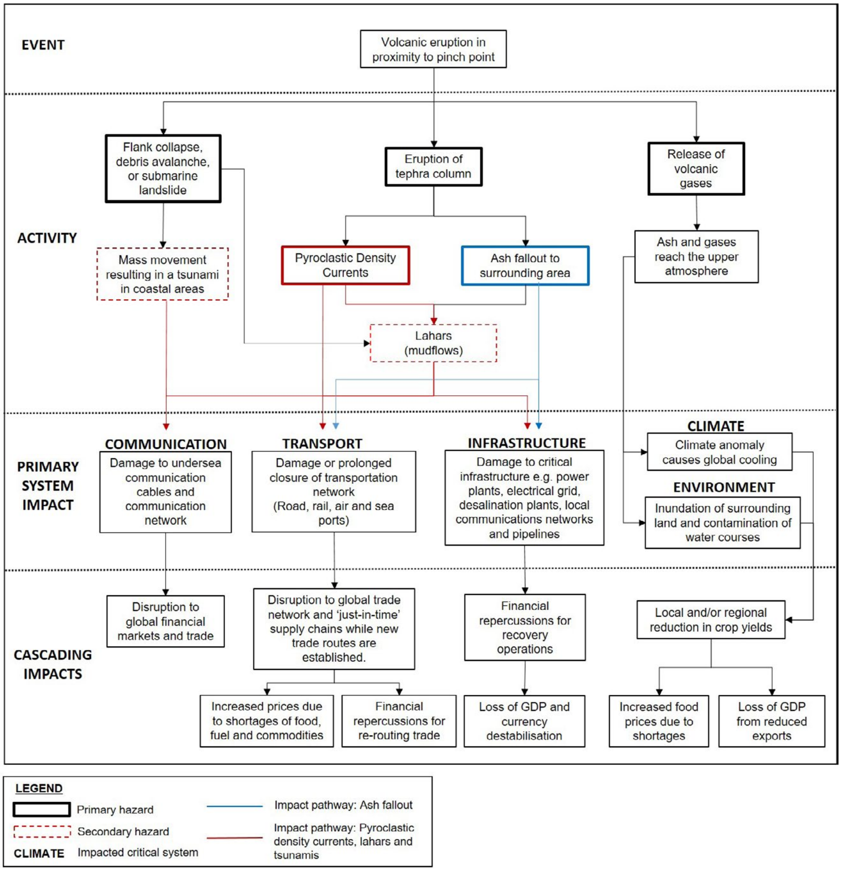

EVENT
Volcanic eruption in proximity to pinch point
ACTIVITY
Flank collapse, debris avalanche, or submarine landslide
Eruption of tephra column
Release of volcanic gases
Mass movement resulting in a tsunami in coastal areas
Pyroclastic Density Currents
Ash fallout to surrounding area
Ash and gases reach the upper atmosphere
Lahars (mudflows)
PRIMARY SYSTEM IMPACT
COMMUNICATION
Damage to undersea communication cables and communication network
TRANSPORT
Damage or prolonged closure of transportation network (Road, rail, air and sea ports)
INFRASTRUCTURE
Damage to critical infrastructure e.g. power plants, electrical grid, desalination plants, local communications networks and pipelines
CLIMATE
Climate anomaly causes global cooling
ENVIRONMENT
Inundation of surrounding land and contamination of water courses
CASCADING IMPACTS
Disruption to global financial markets and trade
Disruption to global trade network and 'just-in-time' supply chains while new trade routes are established.
Financial repercussions for recovery operations
Local and/or regional reduction in crop yields
Increased prices due to shortages of food, fuel and commodities
Financial repercussions for re-routing trade
Loss of GDP and currency destabilisation
Increased food prices due to shortages
Loss of GDP from reduced exports
LEGEND
Primary hazard
Secondary hazard
CLIMATE Impacted critical system
Impact pathway: Ash fallout
Impact pathway: Pyroclastic density currents, lahars and tsunamis

Fig. 1: Cascading system failures from moderate volcanic eruptions. Event tree of lower-magnitude (Volcanic Explosivity Index 3–6) volcanic eruptions in proximity to global critical systems. The event tree demonstrates the propagation of cascading failures of related and interlinked critical systems, due to various eruptive hazards, such as the eruption of tephra column, ash fallout, and pyroclastic density currents, as discussed in this Comment. The figure identifies the pathways from systems that are directly vulnerable to such activities to secondary and tertiary knock-on ramifications for interlinked systems. The blue boxes and arrows depict the impact pathway linked to volcanic ash fallout, whilst the red boxes and arrows show the impact pathway for Pyroclastic Density Currents (PDCs), lahars, and tsunamis. Thick black outlined boxes present primary hazards, whilst dashed outline boxes identify secondary hazards.

We saw an example of the VEI-GCR asymmetry mechanism in play during the 2010 VEI 4 eruption of Eyjafjallajökull, Iceland, whereby a moderate-scale volcanic eruption occurred in proximity to a pinch point of critical systems and networks, resulting in global-scale impacts. During the explosive phase of the event, plumes of volcanic ash were transported on north-westerly winds towards continental Europe,[6] resulting in closure of European airspace, at a loss of US $5 billion to the global economy.[7] This eruption remains the most costly volcanic eruption ever recorded, even when compared to the VEI 6 1991 eruption of Mount Pinatubo, which was the second largest eruption (in terms of tephra ejected) in the last century. The Mount Pinatubo eruption, by contrast, resulted in economic impacts around US $374 million (US $740 million in 2021, recalculated for inflation), despite the eruption being 100 times greater in scale. However, increased globalisation and demand for vital commodities that sustain our societies increased the criticality of the trade and transport networks disabled by the Eyjafjallajökull eruption, driving the global economic impacts and demonstrating an imbalance in humanity's relationship with volcanoes, towards VEI-GCR asymmetry.

Currently, there is little consideration within existing literature of the interplay between critical systems and lower magnitude volcanic activity (VEI 3–6) that mark the new GCR geography, with only a few recent studies mentioning this significant link at all.[8] Where reference is made, it typically focuses on the larger-scale eruption scenarios (VEI 7 and above), and their direct impacts, such as loss of life and damage to infrastructures.[9] However, these studies fail to extend the risk assessments further to consider cascading failure mechanisms that can catapult local systems failures to GCR.[10] Figure 1 illustrates potential cascading system failures with global ramifications that could result from moderate volcanic eruptions of VEI of 3 to 6.

2. Seven Global Pinch Points

On the consideration of an emerging, asymmetric volcanic risk landscape, we highlight at least seven geographical locations, or pinch points, where a convergence of one or more of critical systems occurs, and delineate the particular GCR mechanism each might provoke. These seven pinch points identify localities where we perceive the highest levels of criticality for the global systems and infrastructures they encompass, (e.g. shipping passages with high traffic volumes that cannot be easily re-routed).

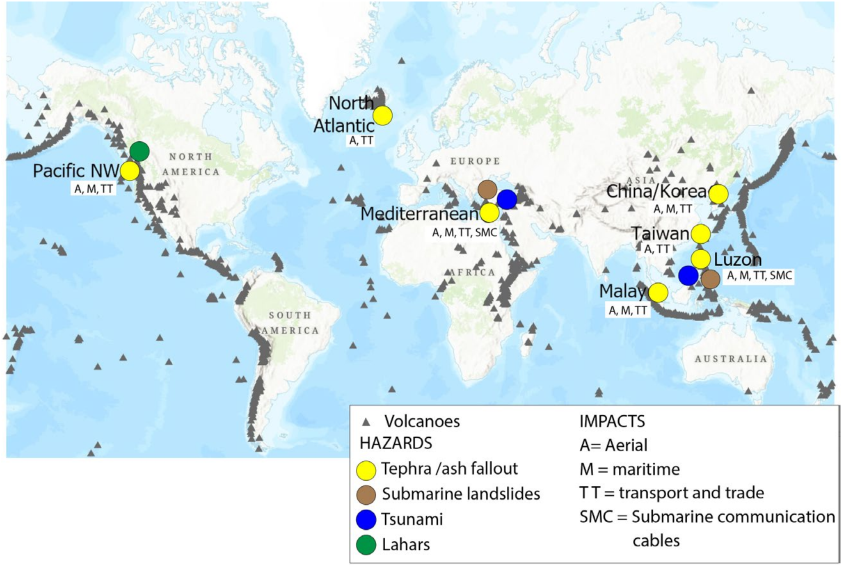


Fig. 2: Seven global pinch points. Map of regions — or pinch points — where clustering of critical systems and infrastructures converge with regions of lower-magnitude volcanic activity (Volcanic Explosivity Index 3–6). These pinch points are presented with the likely associated volcanic hazard activities in circles; where yellow is tephra/ash fallout, brown is submarine landslides, blue are tsunamis, and green are lahars. Each pinch point also includes the potentially impacted systems, including aerial (A), maritime (M), trade and transportation networks (TT), and submarine cables (SMC).

Taiwanese pinch point. The Tatun Volcanic Group (TVG) lies on the northern tip of Taiwan and on the edge of metropolitan Taipei. This volcanic complex was historically active between 2.8 and 0.2 Ma; however, new evidence suggest that it has remained active, with frequent episodes of volcanic-tectonic earthquakes.[11] Taiwan is home for the main manufacturing centre of TSMC, the leading producers of over 90% of the most advanced chips and nodes (equivalent to US $18.9 billion market share)[12] and principal suppliers to the global technology and car industries. An explosive volcanic eruption at TVG could blanket the area in thick tephra deposits, forcing closure of transportation networks, including the Port of Taipei, essential to TSMC's supply chain. Prolonged rupture of critical infrastructures such as the electrical grid that supplies the TSMC manufacturing plants could also cause grave disruption to the global supply of chips and nodes, with severe knock-on implications to the global technology industry and global financial markets.

Chinese-Korean pinch point. The Changbaishan volcanic complex encompassing Mount Paektu straddles the Chinese-North-Korean border, and is most known for its 946 CE "millennium eruption" which was estimated to be a VEI 7 eruption. Tephra deposits from this eruption have been documented as far as Hokkaido, Japan,[13] demonstrating the capability of this volcano to cause widespread disruption in the region. An eruption column, even from a smaller-scale eruption (VEI 4 to 6) at Mount Paektu could be capable of producing a tephra column that would disrupt some of the busiest air routes in the world, such as Seoul to Osaka and Seoul to Tokyo[14] and to maritime traffic traversing the Sea of Japan.

Luzon pinch point. The Luzon Strait is a key shipping passage connecting the South China Sea to the Philippine Sea, and a key route for submarine cables, with at least 17 cables connecting China, Hong Kong, Taiwan, Japan and South Korea. The Luzon Volcanic Arc (LVA) encompassing Mount Mayon, Mount Pinatubo, Babuyan Claro and Taal volcanoes, among others, presents a possible location for an explosive eruption to disrupt the Strait. Volcanic ash and volcanically induced submarine landslides and tsunamis in this region (particularly from submarine volcanic centres) would pose a risk to submarine cable infrastructure within the Strait, and result in closure of the shipping passage. The 2006 7.0 Mw Hengchun earthquake off the south-west coast of Taiwan triggered submarine landslides which severed nine submarine cables in the Strait of Luzon which connect Hong Kong, China, Taiwan, the Philippines and Japan, resulting in near-total internet outages and severely disabling communication capacities (up to 80% in Hong Kong), with knock-on widespread disruptions to global financial markets. These disruptions continued for weeks in the aftermath, with repairs to the cables taking eleven ships 49 days to restore.[15]

Malay pinch point. The Strait of Malacca is one of the busiest shipping passages in the world, with 40% of global trade traversing the narrow route each year.[16] Kuala Lumpur and Singapore both border the Strait and comprise busy aerial and maritime travel and trade hubs. The region is also one of the busiest airspaces in the world, with the aerial route between both cities alone comprising over 5.5 million seats per year.[17] This region is also known to be highly volcanically active, with numerous volcanic centres present along the Indonesian archipelago, such as Mount Sinabung (VEI 4) and Mount Toba in Sumatra, and Mount Merapi (VEI 4) in Central Java. Rupture or either aerial or maritime transportation as a result of a tephra column, could result in severe delays and disruption to global trade. Modelling for a for a VEI 6 eruption at Mount Merapi which only considered the cost of disruption to aerial routes, with the closure of airspace across Malaysia, Indonesia and Singapore, estimated a potential losses of up to US $2.51 trillion dollars of global GDP output loss over a five-year period.[18]

Mediterranean pinch point. Similar to the Straits of Malacca, the Mediterranean is a vital passage for the maritime transportation of goods and commodities from the Middle East and Asia to Europe, and hosts a large network of submarine communications cables connecting Europe to Africa, North America, the Middle East and Asia. A volcanically induced tsunami from a volcanic centre such as Santorini (as happened during the Minoan eruption 3500 BCE), could cause widespread damage to submarine cables and disruption to port facilities and global shipping passages, such as the Suez Canal. The criticality of the Suez Canal was highlighted by the closure of the passage as a result of the stranding of a container ship in March 2021. The six-day closure is estimated to have cost between US $6–10 billion a week to global trade, through delays in cargo transportation and diversion of ships away from the canal.[19] Numerous volcanic centres in the region able to produce such activity, including Mount Vesuvius, Santorini and Campi Flegrei, which are all capable of explosive eruption of VEI 3–6. Additionally, any tephra column produced during an eruption would result in a provisional closure of European airspace, within widespread delays to aerial transport and trade networks.

North Atlantic pinch point. The aerial traffic between London and New York comprises over three million seats per year.[20] Disruption to this critical artery could cause widespread disruption and delay to global trade and transportation networks. Volcanic centres in Iceland are a potential source for this disruption, with numerous volcanic centres producing explosive events of VEI 3–6, including Katla (1918), Hekla (1947) and Grímsvötn (2011).

Pacific Northwest pinch point. An eruption of a Cascades volcano, such as Mount Rainier, Glacier Peak or Mount Baker in Washington, would have the potential to trigger mass flows, such as debris avalanches or lahars, resulting from the melting of glaciers and ice caps, with the potential to reach Seattle.[21] The Osceola mudflow generated around 5,600 years ago at Mount Rainier travelled over 60 miles to reach Puget Sound at the site of the present-day Port of Tacoma, Seattle. The generation of a similar-scale mass flow, and combined with any ash fall towards Seattle, would force provisional closure of airports and seaports, which account for 2.5% of the US's total traffic respectively.[22] Volcanic ash might also affect a wider airspace including parts of Canada (e.g. Vancouver) and US cities such as Portland. Scenario modelling for a VEI 6 eruption at Mount Rainier with volcanic ash closing airspace across the northern USA and parts of Canada predict potential losses of up to US $7.63 trillion dollars of global GDP output loss over a five-year period.[23]

3. Reconsidering Volcanic Risk Assessments

By converging critical systems within pinch point localities and placing them at the interface with regions of potential volcanic activity, we have manufactured a new type of GCR from lower VEI 3 to 6 magnitude eruptions; a narrative that has previously been neglected by the volcanic risk community. The identification of "pinch points" tilts the relationship between volcanic activity and GCRs, towards VEI-GCR asymmetry, thereby presenting a current gap in our approach to volcanic risk assessment, and disaster prevention and mitigation practices. We suggest that the community should now consider this risk asymmetry in assessments, and work to fully understand the systemic vulnerabilities that may catapult us from a moderate magnitude volcanic eruption (VEI 3 to 6) to a GCR.

As preparedness measures in the pre-disaster phase, we propose that systems mapping and evidence-based foresight activities — such as horizon scanning and event tree analysis — be more systematically incorporated into work to identify the full extent and nature of our VEI-GCR asymmetry, and identify opportunities where resilience can be built towards Global Catastrophic Volcanic Risk. These activities ought to rely on expert elicitation, including from natural and geophysical sciences, civil engineering, and economics. The asymmetry mechanism discussed here in the context of volcanic hazards, is also likely applicable to other geophysical phenomena; a similar approach could be considered for seismic, hydrogeological, and meteorological hazards alike, where this is not already the case.

Unlike super-volcanic eruption scenarios where we have little opportunity for prevention, we can work to reduce the fragility and exposure of our critical systems to rapid onset natural events, and ultimately increase our resilience to GCRs.

Acknowledgements

Lara Mani is supported by a grant from Templeton World Charity Foundation, Inc. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation, Inc.

Notes and References

[1] Papale, Paolo and Warner Marzocchi. 'Volcanic threats to global society', *Science, 363*(6433) (22 March 2019): 1275. https://doi.org/10.1126/science.aaw7201; Rampino, Michael R. 'Supereruptions as a threat to civilizations on Earth-like planets', *Icarus, 156*(2) (1 April 2002): 562–69. https://doi.org/10.1006/icar.2001.6808

[2] Newhall, Chris, Stephen Self and Alan Robock. 'Anticipating future Volcanic Explosivity Index (VEI): 7 eruptions and their chilling impacts', *Geosphere, 14*(2) (1 April 2018): 572–603. https://doi.org/10.1130/GES01513.1

[3] Bostrom, Nick and Milan M. Ćirković. *Global Catastrophic Risks*. Oxford University Press (2011).

[4] Oppenheimer, Clive. *Eruptions That Shook the World*. Cambridge University Press (2011). https://doi.org/10.1017/CBO9780511978012; Stothers, Richard B. 'The great Tambora eruption in 1815 and its aftermath', *Science, 224*(4654) (15 June 1984): 1191. https://doi.org/10.1126/science.224.4654.1191

[5] Avin, Shahar et al. 'Classifying global catastrophic risks', *Futures, 102* (1 September 2018): 20–26. https://doi.org/10.1016/j.futures.2018.02.001; Hinchey, M. and L. Coyle. 'Evolving critical systems: A research agenda for computer-based systems', *2010 17th IEEE International Conference and Workshops on Engineering of Computer Based Systems* (2010): 430–35. https://doi.org/10.1109/ECBS.2010.56

[6] Gudmundsson, Magnús T. et al. 'Ash generation and distribution from the April-May 2010 eruption of Eyjafjallajökull, Iceland', *Scientific Reports, 2*(1) (14 August 2012): 572. https://doi.org/10.1038/srep00572

[7] Oxford Economics. 'The economic impacts of air travel restrictions due to volcanic ash', *Oxford Economics* (1 May 2010). https://www.oxfordeconomics.com/my-oxford/projects/129051

[8] Papale and Marzocchi (2018); Newhall, Self, and Robock (2018).

[9] Rampino (2002); Oppenheimer (2011); Wilson, G. et al. 'Volcanic hazard impacts to critical infrastructure: A review', *Journal of Volcanology and Geothermal Research, 286* (1 October 2014): 148–82. https://doi.org/10.1016/j.jvolgeores.2014.08.030

[10] Wilson et al. (2014).

[11] Pu, H. C. et al. 'Active volcanism revealed from a seismicity conduit in the long-resting Tatun volcano group of Northern Taiwan', *Scientific Reports, 10*(1) (9 April 2020): 6153. https://doi.org/10.1038/s41598-020-63270-7

[12] Hille, Kathrin. 'TSMC: How a Taiwanese chipmaker became a linchpin of the global economy', *Financial Times* (24 March 2021). https://www.ft.com/content/05206915-fd73-4a3a-92a5-6760ce965bd9

[13] Machida, Hiroshi and Fusao Arai. 'Extensive ash falls in and around the Sea of Japan from large late Quaternary eruptions', *Journal of Volcanology and Geothermal Research, 18*(1) (1 October 1983): 151–64. https://doi.org/10.1016/0377-0273(83)90007-0

[14] OAG Aviation Worldwide Limited. 'Busiest routes 2020', *OAG Free Reports* (April 2020). https://www.oag.com/hubfs/free-reports/2020-reports/busiest-routes-2020/busiest-routes-2020.pdf?hsCtaTracking=9a937560-d748-4f4f-bb61-3f5063040294%7Cd74a14a5-13fb-4a03-9c32-ec7825bd0d91

[15] Sunak, Rishi. 'Undersea cables: Indispensable, insecure', *Policy Exchange* (2017). https://policyexchange.org.uk/wp-content/uploads/2017/11/Undersea-Cables.pdf

[16] Bailey, Rob and Laura Wellesley. 'Chokepoints and vulnerabilities in global food trade', *Chatham House* (27 June 2017). https://www.chathamhouse.org/sites/default/files/publications/research/2017-06-27-chokepoints-vulnerabilities-global-food-trade-bailey-wellesley-final.pdf

[17] OAG Aviation Worldwide Limited (2020).

[18] Mahalingam, A., A. Coburn, C. J. Jung, J. Z. Yeo, G. Cooper and T. Evan. *Impacts of Severe Natural Catastrophes on Financial Markets*. Cambridge Centre for Risk Studies (2018).

[19] Russon, Mary-Ann. 'The cost of the Suez Canal blockage', *BBC News* (29 March 2021). https://www.bbc.com/news/business-56559073

[20] OAG Aviation Worldwide Limited (2020).

[21] Vallance, James W. and Kevin M. Scott. 'The Osceola mudflow from Mount Rainier: Sedimentology and hazard implications of a huge clay-rich debris flow', *GSA Bulletin, 109*(2) (1 February 1997): 143–63. https://doi.org/10.1130/0016-7606(1997)109<0143:TOMFMR>2.3.CO;2

[22] Mahalingam et al. (2018).

[23] Mahalingam et al. (2018).

13. Re-Framing the Threat of Global Warming: An Empirical Causal Loop Diagram of Climate Change, Food Insecurity and Societal Collapse

C. E. Richards, R. C. Lupton and J. M. Allwood

Highlights:

- Understanding the existential threat of climate change is essential for good risk management; however, our knowledge of the pathways through which climate change could cause societal collapse is underdeveloped. This chapter aims to identify and structure an empirical evidence base of the climate change, food insecurity, and societal collapse pathway.
- The authors first review the societal collapse and existential risk literature and define a set of determinants of societal collapse. They then develop an original methodology, using these determinants as societal collapse proxies, to identify an empirical evidence base of climate change, food insecurity, and societal collapse and structure this using a novel-format Causal Loop Diagram (CLD).
- The resulting evidence base varies in temporal and spatial distribution of study and in the type of data-driven methods used. For example, the link between *food insecurity* and

https://doi.org/10.11647/OBP.0360.13

conflict was found to have been investigated mostly by statistical analyses, whereas the links between *food insecurity* and *migration*, and *food insecurity* and *natural mortality* were investigated mostly by interviews and surveys.

- The CLD documents the spread of the evidence base and enables exploration of how the effects of climate change may undermine agricultural systems and disrupt food supply, which can lead to economic shocks and socio-political instability, as well as starvation, migration and conflict.
- Suggestions are made for future work that could build on this chapter to further develop our qualitative understanding of, and quantitative complex systems modelling capabilities for analysing, the causal pathways between climate change and societal collapse. In particular, it highlights important factors at global scale and national granularity, such as the geographic distribution of population and natural resources, international interactions (such as food trade, conflict and migration), and institutional breakdown.

This chapter uses a systems dynamics approach and proposes that a Causal Loop Diagram can be utilised in order to construct an evidence base of the relationships between climate change, food insecurity and societal collapse. Novel approaches to investigating and mapping systemic and complex risk interactions are also explored in chapters throughout this volume, including Chapter 20.

1. Introduction

Despite recent social protests and climate emergency declarations, efforts to mitigate climate change to date are insufficient.[1] Greenhouse gas (GHG) emissions continue to rise and global warming above 3 °C is increasingly likely this century.[2] There is emerging evidence of amplifying feedbacks accelerating[3] and dampening feedbacks decelerating.[4] These feedbacks exacerbate the possibility of runaway global warming,[5] estimated at 8 °C or greater by 2100.[6] Such temperature increases translate to a range of real dangers,[7] shifting the narrow climate niche within which humans have resided for millennia.[8]

Looking beyond the framing of "global warming", there is concern that the effects of climate change may pose an existential risk to humanity, one that threatens "societal collapse" or even extinction.[9] Understanding these worst-case scenarios is essential for good risk management.[10] Improving awareness of potential pathways through which climate change poses such a risk can help inform decision-making about interventions.[11] Considering societal impacts that are more tangible for individuals, businesses and governments,[12] and better aligned with conventional risk priorities,[13] may facilitate more effective action to mitigate climate change.[14]

A number of pathways through which climate change could cause societal collapse have been identified, one being via food insecurity.[15] Climate change is predicted to undermine agricultural systems and disrupt food supply,[16] which may lead to economic shocks, socio-political instability as well as starvation, migration and conflict at local through to global scale.[17] While the climate science underpinning global warming estimates is well established,[18] albeit subject to sensitivities, the uncertainties increase significantly when we start to consider these tangible societal impacts given the complex relationships involved.[19] Our understanding of worst-case scenarios, and particularly of empirical evidence addressing the causal pathways through which climate change may cause societal collapse, is underdeveloped.[20]

In this chapter we aim to identify and structure an empirical evidence base of the relationships between climate change, food insecurity and societal collapse. We do this using Causal Loop Diagrams (CLD), a system dynamics approach that is useful for visualising the relationships between variables in a complex system.[21] This chapter is organised as follows. In Section 2, we review the societal collapse and existential risk literature to refine the aim introduced above. In Section 3, we develop an original methodology to establish a new empirical evidence base and create a novel-format CLD of causal pathways between climate change and societal collapse. In Section 4, we present and discuss the results from the application of this methodology to the climate change, food insecurity and societal collapse causal pathway of interest. We conclude, in Section 5, by identifying avenues of future work that may build upon this chapter.

2. Literature Review

To refine the aim of this chapter, introduced in Section 1, our review examines whether there is historical evidence of climate change as a mechanism of societal collapse and to what extent have causal pathways been documented to inform our understanding of climate change as an existential threat to contemporary society.

We first define the terms "existential risk" and "societal collapse" as used in this chapter. Adopting Ord's definition, "an existential risk is a risk that threatens the destruction of humanity's long-term potential" be it incomplete destruction, such as societal collapse, or complete destruction, such as extinction.[22] Adopting Kemp's definition, societal collapse is an "enduring loss of population, identity [and/or institutional] complexity";[23] it may be abrupt or gradual, but is typically rapid because it is notably transformative, and may be experienced by a local, national or the global community of people. Fig. 1 presents a conceptual model of societal collapse, synthesised from the broader literature, to provide further contextual definition.

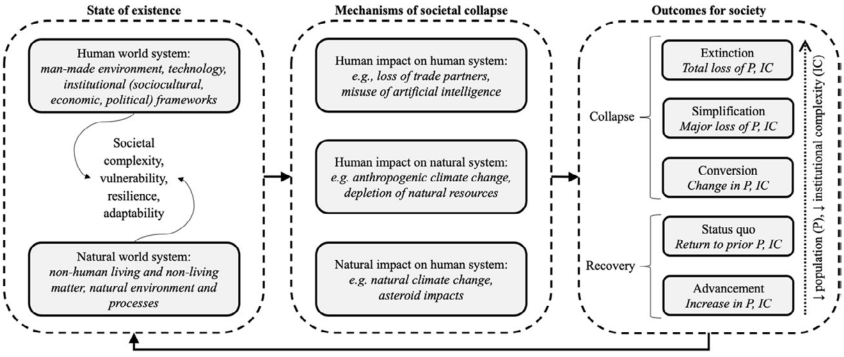

State of existence
Human world system: man-made environment, technology, institutional (sociocultural, economic, political) frameworks
Societal complexity, vulnerability, resilience, adaptability
Natural world system: non-human living and non-living matter, natural environment and processes
Mechanisms of societal collapse
Human impact on human system: e.g., loss of trade partners, misuse of artificial intelligence
Human impact on natural system: e.g. anthropogenic climate change, depletion of natural resources
Natural impact on human system: e.g. natural climate change, asteroid impacts
Outcomes for society
Collapse
Extinction Total loss of P, IC
Simplification Major loss of P, IC
Conversion Change in P, IC
Recovery
Status quo Return to prior P, IC
Advancement Increase in P, IC
↓ population (P), ↓ institutional complexity (IC)

Fig. 1: Conceptual model of the overarching process of societal collapse.

The rise and fall of civilisations has been documented since the earliest recordings of history and is increasingly studied to inform our understanding of societal collapse.[24] We consider two types of historical studies that provide insight into climate change as a mechanism of societal collapse in the past. We note that other mechanisms are also discussed in the literature, and there is debate about the role of different mechanisms in particular societal collapse events.

The first type of historical study empirically investigates an individual societal collapse event using primary sources, including anthropological, archaeological and paleontological data. Based on such data analysis, natural climate change has been asserted as a mechanism of societal collapse in many of these case studies, as established by de Menocal[25] and Weiss and Bradley.[26] For example, Hodell et al.,[27] Haug et al.,[28] and Medina-Elizalde and Rohling[29] analyse paleoclimate data alongside the archaeological record to show that drought conditions, driven by climate change likely due to solar forcing, contributed to the collapse of the Classic Maya civilisation of Mesoamerica in ~8–10th century CE. Weiss et al.,[30] Cullen et al.,[31] and Cookson et al.[32] show that regional aridity, driven by climate change likely due to volcanic forcing, contributed to the collapse of multiple societies across Mesopotamia, including the Akkadian Empire in ~22nd century BCE. Similarly, natural climate change has been implicated in the collapse of multiple Late Bronze Age societies around the Mediterranean,[33] including Mycenaean Kingdoms in ~12th century BCE,[34] the Harappan Civilization of South Asia in ~19th century BCE,[35] the Angkor Empire of Southeast Asia in ~15th century CE,[36] multiple Chinese Dynasties[37] and civilisations along the Silk Road[38] during the previous millennium, the Norse Vikings of Greenland in ~16th century AD,[39] and the Tiwanaku Empire of Pre-Columbian South America in ~10th century CE[40] amongst others.

This first type of studies establishes precedence of natural climate change as a mechanism of societal collapse throughout history, demonstrating the risk that anthropogenic climate change similarly poses to contemporary society. However, the events examined occurred more than 100 years ago, with most dating back to ancient history, when societies were relatively isolated. Because these case studies predate contemporary society, they do not provide empirical evidence of anthropogenic climate change in context of today's highly interconnected society.[41]

Statistical evaluation of the frequency and significance of natural climate change relative to other mechanisms of societal collapse identified across these case studies has not yet been established within the literature. However, the second type of historical study does qualitatively examine collections of these case studies to develop theories of predominant modes of societal collapse. Three major modes are observed, as follows.

Fagan[42] and McMichael[43] focus on natural impact on the human system across multiple civilisations, concluding that natural climate change is predominant having significantly influenced human existence throughout history. Over the past 12,000 years, the natural and human systems developed within the stable climate niche of the Holocene Epoch.[44] The associated geographic endowments governed human transition from band societies based on foraging to complex societies based on agriculture. Unfavourable subtle (e.g. weather variations) and drastic (e.g. natural disasters) shifts in climate influenced the collapse of complex societies either by direct loss of life or indirectly via resource insecurity. In particular, in this mode, typically, the loss of agriculture led to de-population via famine, migration or conflict due to food insecurity.

Ponting,[45] Wright,[46] and Diamond[47] focus on human impact on the natural system across multiple civilisations, concluding that human overpopulation and overexploitation relative to the carrying capacity of the environment is predominant. Societal collapse via environmental degradation often involved unsustainable agriculture, exacerbated by natural climate change, leading to de-population as well as institutional breakdown via loss of economic stability and socio-political dysfunction due to magnified inequality. This mode aligns with early "Malthusian catastrophe",[48] "tragedy of the commons",[49] and "overshoot-and-collapse"[50] theories.

In their 12-volume *magnum opus* exploring the rise and fall of 28 civilizations, Toynbee concludes that "great civilizations are not murdered [but rather] they take their own lives."[51] Building on this, Tainter,[52] Acemoglu and Robinson,[53] and Johnson[54] focus on human impact on the human system across multiple civilisations, concluding that societal complexity in relation to problem-solving inability (e.g. environmental degradation) and institutional dysfunction (e.g. inequality and oligarchy internally, trade ally and hostile neighbour relations externally) is predominant. As a society becomes more complex, it reaches a point beyond which "continued investment in complexity as a problem-solving strategy yields a declining marginal return" and it will be at risk of collapsing under its own weight via institutional breakdown and de-population.[55] This mode aligns with ″energy returned on energy invested″ theory,[56] applied to explore societal collapse by Homer-Dixon.[57]

Diamond,[58] Turchin,[59] and Schwartz and Nichols[60] examine why some civilisations have been able to thrive or recover, rather than collapse. They similarly conclude that societies have flourished due to combinations of favourable geographic endowment, managing their existence within the carrying capacity of the natural system, and co-operative action in problem-solving.

This second type of studies highlights that societal collapse involves a complex nexus of factors and dynamically interlinked events. For instance, Gibbon details how all three of the modes, described in the preceding five paragraphs, contributed to the collapse of the Roman Empire.[61] These modes of societal collapse, although based on empirical evidence pre-dating contemporary society, describe key aspects of the anthropogenic climate change problem faced today. While these studies describe causal pathways of relevance, to the best of our knowledge, no study has used CLDs to untangle the complexity and give structure to the dense information in this evidence base.

Across these historical studies, we observe no apparent temporal or spatial influence on the occurrence of societal collapse. Rather, societal collapse has been described as occurring in various forms, whether it be by known “white-swan” or surprise “black swan” events,[62] in different geographic locations and times throughout history. Additionally, a quantitative statistical analysis by Arbesman shows that societal collapse has occurred randomly and independent of civilisation life-spans.[63] These qualitative and quantitative observations highlight that any society may be susceptible to collapse, much in-line with the *Red Queen Hypothesis* of the *Law of Extinction*.[64]

From these historical studies, we observe sets of secondary determinants for each of the primary determinants introduced in Fig. 1, which are defined in Fig. 2. Considering a geographically bounded society, *emigration* refers to any permanent departure of population including both voluntary or forced migration, *conflict mortality* accounts for deaths directly arising from any form of domestic or international conflict (e.g. due to war), and *natural mortality* accounts for deaths related to domestic environmental conditions (e.g. due to famine). The *loss of socio-cultural norms, political structures* or *economic value* accounts for that which notably transforms the identity and institutions of the society.

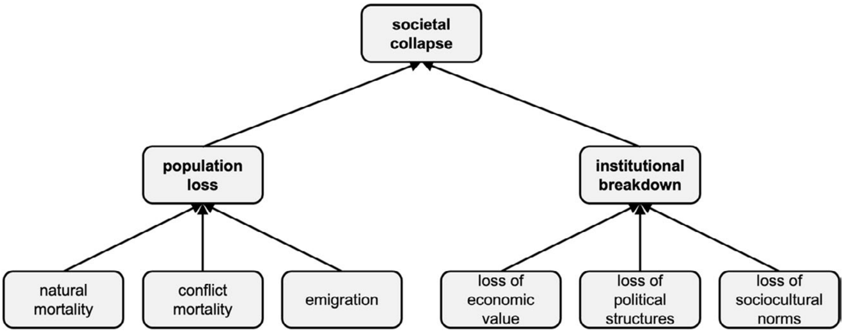


Fig. 2: Primary and secondary determinants of societal collapse for a bounded society.

In addition to these historical studies, we consider the relatively nascent studies of existential risks (X-risks) that provide insight into how climate change may trigger societal collapse in the future.

Comprehensive surveys of X-risks reveal mechanisms that could cause the collapse of contemporary society. Bostrom and Ćirković,[65] Rees,[66] and Ord[67] provide eminent scholarly treatment of the field, drawing from the academic literature. The World Economic Forum[68] and Global Challenges Foundation[69] produce global risk reports drawing from decision-makers and experts across intergovernmental and non-governmental organisations. These surveys establish that many historically observed mechanisms of societal collapse, including natural climate change, remain applicable as X-risks today. However, the state of existence of contemporary society has led to a different landscape in which these mechanisms apply, and to a number of unprecedented mechanisms, including anthropogenic climate change. Ehrlich and Ehrlich[70] and Häggström[71] note that although increased complexity, such as globalisation and technological advancement, can increase a society's resilience and adaptability, it can also increase vulnerability. For example, globalisation increases resilience to local agricultural production shocks through access to global markets; however, it also increases vulnerability through exposure to sudden reversal in connectivity, such as trade restrictions.[72] Some geoengineering technologies, for example, may enable society to mitigate and adapt to climate change; however, they may also increase vulnerability to termination shocks, where failure of the technology exposes society to sudden temperature increases.[73] In this highly interconnected landscape, "synchronous"[74] and "cascading"[75] failures create the potential for mechanisms and outcomes of societal collapse, once contained to a single localised civilisation, to rapidly spread across multiple nations and impact humanity on a global scale.

Works by Lynas,[76] Wallace-Wells[77] and Gowdy[78] draw on the scientific climate change literature to explore hypothetical futures under best- to worst-case scenarios. The scenarios consider the feedbacks within the natural system that could worsen, as well as the potential for humans to mitigate, anthropogenic climate change. Shifts in average weather (e.g. temperature) and natural disasters (e.g. floods) affected by climate change could impact human mortality directly. These two effects, coupled with sea level rise due to melting of ice caps, could indirectly impact human mortality via degradation of the natural world system (e.g. land quality) and the human world system (e.g. infrastructure failures) resulting in resource and service insecurity. This insecurity could impact institutional stability, resulting in economic loss, political dysfunction and social unrest, as well as migration and conflict. The hypothetical outcomes for contemporary society against the threat of anthropogenic climate change range from dystopian (collapse) to utopian (recovery).

These futures studies identify endpoints of different causal pathways between anthropogenic climate change effects and potential impacts on the human world system, with the latter reflecting key determinants of societal collapse observed in the historical studies. Scholars have made limited in-roads to empirically investigating the top-level relationships between some of these endpoints using recent datasets. The direct links between climate change and the endpoint impacts of mortality, conflict and migration are, respectively, examined by Mora et al.,[79] Hsiang et al.[80] and Hauer et al.[81] The feedback between migration and conflict driven by climate change is examined by Abel et al.[82] The direct links between climate change and the endpoint impacts of economic loss, political instability and shifts in cultural norms are examined by Burke et al.,[83] Sofuoğlu and Ay,[84] and Adger et al.[85] respectively. However, the complex bottom-level links between and surrounding these endpoints are generally ill understood,[86] and the strength of empirical evidence is poorly documented from a systems science perspective.[87] To the best of our knowledge, no study has empirically examined how the impacts of climate change could explicitly translate into societal collapse for contemporary society. We do not have a clear picture of climate change as a systemic risk to our globalised society, particularly at spatial scales accounting for the heterogeneity of individual identity, business governance and policymaking across nations, and international exchanges. This limits our ability to understand feedbacks, identify intervention points, develop quantitative models and inform strategies to minimise the risk of societal collapse occurring in the future.[88]

Given the insights from this review, we refine the aim of this chapter as follows. Firstly, the empirical evidence base should specifically address contemporary society. Secondly, the CLD should be constructed at a scale and granularity that addresses the heterogenous characteristics of nations and international interactions. The refined aim of this chapter is thus to identify an empirical evidence base of climate change, food insecurity and societal collapse in contemporary society and structure the evidence base with a CLD defined at global scale and national granularity.

3. Methodology

A two-stage framework, consisting of five steps, was developed to achieve the aim of this chapter. For each step, below, we first introduce it generically and then describe its application to our specific analysis of the climate change, food insecurity and societal collapse causal pathway.

3.1. Stage 1: Establishing an empirical evidence base of societal collapse in contemporary society

Step I deploys societal collapse proxies via a key word search to identify "evidence points", which in this instance may be considered data points, in the form of publications that empirically examine the causal pathway of interest in contemporary society.

The determinants defined in Fig. 2 provide these societal collapse proxies to establish the new empirical evidence base in lieu of historical societal collapse events pre-dating contemporary society. The *population loss* set are straightforward to isolate, consistent to measure across nations and describe tangible consequences. The *institutional breakdown* set are relatively less so. Thus, the societal collapse proxies adopted in this study were *natural mortality* (i.e. *starvation*, with respect to food insecurity), *conflict mortality* and *emigration*; subsequent studies could use the *institutional breakdown* set. Key words were selected based on terminology of climate change, food insecurity and the societal collapse proxies. Peer-reviewed journal articles were chosen as the form of evidence point in this study; subsequent studies could use other publications, such as books and reports.

The keyword search was performed in Scopus. A record of the search is contained in the Supplementary Information (A.). Approximately 3,000 publications were reviewed by reading the title, abstract and main body as needed. Evidence points were selected based on satisfaction of the following criteria: the publication (a) is a peer-reviewed, English-language, journal article; (b) uses empirical, data driven methods; (c) examines the period from 1990 to present (2019), representative of contemporary society; and (d) primarily examines the causal pathway of interest. We made an exception to (a) to include the most recent *Limits to Growth* book,[89] which was not itself a search result but documents the World3 model that was identified in the search results. We note that (b) precluded selection of review or essay-style publications; however, we found that these were often discussed in the literature review of selected evidence points, so were, nonetheless, accounted for indirectly.

This step resulted in a new empirical evidence base consisting of 41 evidence points, which are summarised in Fig. 4.

Step II defines a custom colour-coded typology for the new empirical evidence base. This typology is used in Stage 2, to construct a final CLD (f-CLD) in a novel format showing the spread of the evidence base across the system.

In this study, we were interested in the methodological spread as this provides information on data that may be useful for future studies. Four methodological categories were identified in the new empirical evidence base. Each evidence point was classified into one of these categories and assigned a colour coding, namely: quantitative complex systems model — red; statistical analysis of quantitative dataset — blue; collection /

analysis of qualitative interview / survey data — green; quantitative data-led case study / scenario — yellow.

The resulting typology of the new empirical evidence base is shown in Fig. 4.

3.2. Stage 2: Constructing a novel-format causal loop diagram from the empirical evidence base

Step III involves creating an individual CLD (i-CLD) for each evidence point to clearly structure the complex causal relationships examined. These i-CLDs provide the building blocks from which to construct the f-CLD in Step IV.

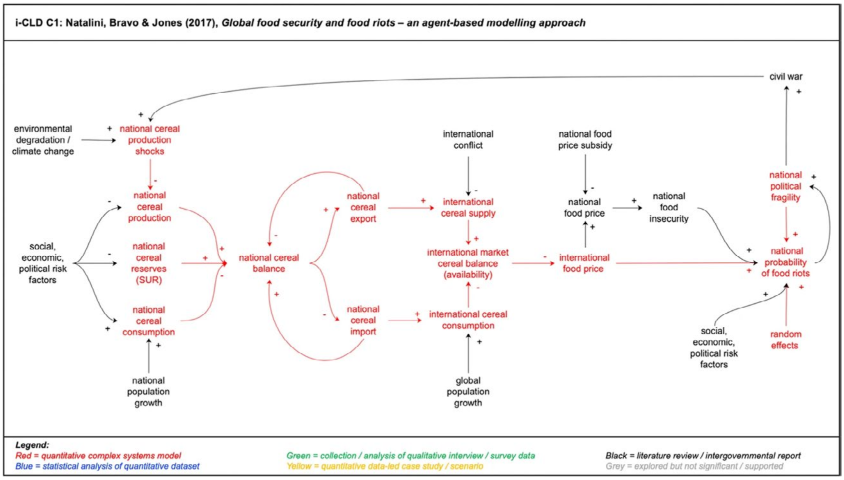

i-CLD C1: Natalini, Bravo & Jones (2017), Global food security and food riots – an agent-based modelling approach
civil war
environmental degradation / climate change
national cereal production shocks
international conflict
national food price subsidy
national political fragility
national cereal production
national cereal export
international cereal supply
national food price
national food insecurity
social, economic, political risk factors
national cereal reserves (SUR)
national cereal balance
international market cereal balance (availability)
international food price
national probability of food riots
national cereal consumption
national cereal import
international cereal consumption
social, economic, political risk factors
random effects
national population growth
global population growth
Legend:
Red = quantitative complex systems model
Blue = statistical analysis of quantitative dataset
Green = collection / analysis of qualitative interview / survey data
Yellow = quantitative data-led case study / scenario
Black = literature review / intergovernmental report
Grey = explored but not significant / supported

Fig. 3: One of the i-CLDs created for each of the 41 evidence points in the new empirical evidence base of climate change, food insecurity and societal collapse in contemporary society.

The process to create an i-CLD is as follows. The corresponding evidence point was examined in its entirety to identify and record key information in the form of variables (nodes), links (arrowed lines) and relationship notation (positive or negative). Key information derived from the original data-driven content, i.e. the main analysis, of the evidence point was colour coded in the i-CLD according to the typology classification established in Step II. Any relationships hypothesised but not supported by the main analysis were coloured grey. Key information derived from other content, i.e. the literature review, of the evidence point was coloured black. The scale and granularity of the i-CLD was recorded as detailed in the evidence point. This process was repeated for each evidence point in isolation until a complete set of i-CLDs was produced for the new empirical evidence base.

All 41 i-CLDs created in this study are contained in the Supplementary Information (B.). One of the i-CLDs is shown in Fig. 3 as an example.

Step IV reconciles the set of i-CLDs into a standardised format in order to construct the f-CLD of the system of interest at the desired scale and granularity.

The standardisation process has two aspects. One aspect is related to component (variables and links) definition, necessary to maximise clarity of the f-CLD while covering all information contained in the evidence base. This addresses the typical challenge of CLDs becoming dense and overcomplicated, which decreases their utility. The other aspect is related to level of aggregation, necessary to ensure the f-CLD conveys information at the intended scale and granularity. The standardisation is an iterative process, as follows.

The ~950 variables from the set of 41 i-CLDs were recorded on a blank worksheet for the f-CLD, without links between them. A clustering approach was used to reconcile these variables into like groups. For each group, an overarching major node was isolated and the i-CLD variables in the group were virtually deposited into a matrix for that major node. For example, *drought, sea level rise* and *crop disease* were some of the i-CLD variables clustered into an *environmental risk factors* f-CLD major node matrix. The f-CLD major nodes were defined at a level of aggregation representative of a nation. Doing so effectively scaled down any global or regional aggregation, and scaled up any sub-national or local aggregation, in the i-CLD variables. For example, *household food imports* was an i-CLD variable of local aggregation that was scaled up to national *food imports* (*trade*) in the f-CLD.

The ~1150 links from the set of 41 i-CLDs were reconciled into arrowed lines between the major nodes in the f-CLD. This sometimes-required interpretation of implied causality in the i-CLD relationships in order to route them across the major nodes in the f-CLD. For example, where an i-CLD showed a direct link from *international food price* to *conflict* variables, this was routed using arrowed lines from *international food price* to national *food price* to *food accessibility* to *food insecurity* and finally to *conflict* major nodes defined in the f-CLD. Where there was a discrepancy between relationship descriptions, the relationship with the most supporting i-CLDs was adopted.

The interim f-CLD produced at the end of each standardisation iteration was examined to determine whether the major node definition could be refined to maximise clarity. For example, in one iteration *water* and *land* were defined as separate major nodes, but examination determined that each had the same arrowed lines to other major nodes; therefore, another iteration was undertaken with *water* and *land* now clustered under a single *natural resources* major node in order to minimise redundant arrowed lines. This process was iterated several times until an f-CLD had been constructed at an appropriate level of detail for this study. Additionally, relevant literature reviewed in Section 2[90] was cross-referenced, but not included as evidence points, to ensure comprehensive coverage of key relationships in the f-CLD.

The standard-format f-CLD, consisting of uncoloured and unweighted components, resulting at the end of this step is contained in the Supplementary Information (C.).

Step V maps each i-CLD to the f-CLD using a weighted (line thickness) typology (colour-coded) approach. This visually documents the spread of the evidence base across the system described by the f-CLD.

The process to map an i-CLD to the f-CLD is as follows. Each variable (node) of the i-CLD was assigned to its corresponding major node(s) in the f-CLD. Each link (arrowed line) of the i-CLD was assigned to a corresponding route along the arrowed lines in the f-CLD. Each time an arrowed line in the f-CLD had an i-CLD link assigned to it, an incremental weighting of one-unit line thickness in the corresponding typology colour-coding of the i-CLD link was added to the f-CLD arrowed line. This process was repeated for each of the 41 i-CLDs until all had been mapped to the f-CLD. A record of this process for each of the 41 i-CLDs is contained in the Supplementary Information (D.).

The novel-format f-CLD, consisting of colour-coded and weighted components, resulting at the end of this final step is presented in Fig. 5.

4. Results and Discussion

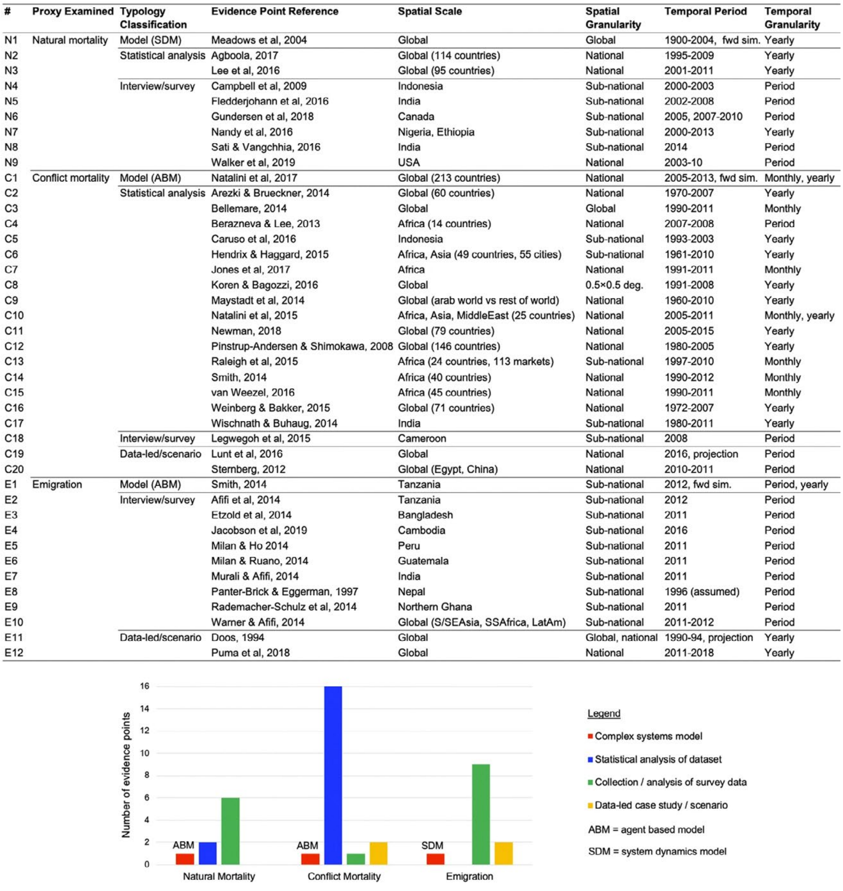
The new empirical evidence base and novel-format CLD of climate change, food insecurity and societal collapse in contemporary society resulting from the application of our original methodology (Section 3) are discussed in turn below.

4.1. Empirical evidence base of climate change, food insecurity and societal collapse in contemporary society

The new empirical evidence base (Section 3, Step I), along with its colour-coded typology (Section 3, Step II), is presented in Fig. 4. It consists of 41 evidence points, of which 9 examine the *natural mortality* (i.e. *starvation*, with respect to food insecurity), 20 the *conflict mortality* and 12 the *emigration* societal collapse proxy, alongside other human and natural world system factors. We discuss three key aspects of the evidence base, namely temporal and spatial distribution, data-driven method distribution, and advantages of each data-driven methods, below.

The temporal scale and granularity of study varies across the evidence base; however, our methodology limited the possible scale of study to the period from 1990 to present, representative of contemporary society. Within this period, approximately half of the evidence points cover a scale of less than one decade and the other half a scale of greater than one decade. Approximately half of the evidence points conduct analyses at yearly granularity and the other half conduct analyses at granularity greater than one year, with only a few studies conducting analyses at monthly granularity. The spatial scale and granularity of study varies across the evidence base. Approximately one third of the evidence points investigate the system at a global scale, with the remaining two thirds focusing on regional or national scales, primarily in Africa as well as the Middle East and Asia. Approximately half of the evidence points analyse the causal pathway at sub-national granularity, with the other half primarily focusing on national-level granularity. This variation provided different coverage of the complex relationships within the system, which was informative for constructing our CLD.

The distribution of data-driven methods used across the evidence base is notably different for each societal collapse proxy. Evidence points for *natural mortality* mostly use collection/analysis of interview/survey data. This is likely because the minimum daily food intake for human survival is well established;[91] as such, statistical analysis of food and mortality data sets would not yield significantly new insights into thresholds whereas interviews/surveys can provide insight into an individual's circumstances influencing this relationship. Evidence points for *conflict mortality* mostly use statistical analysis of existing datasets. This likely reflects the interest in rigorously curated conflict datasets, such as UCDP/PRIO,[92] across the conflict and peace fields. Evidence points for *emigration* mostly use collection/analysis of interview/survey data, likely because this provides nuanced insight into an individual's decision to migrate. It may also be due to data availability and quality challenges that limit quantitative statistical analyses, which are being addressed by groups such as the International Organization for Migration's Global Migration Data Analysis Centre.[93] Amongst these data challenges, it is important to recognise the issue of reconciling different types of voluntary and forced migration with causal drivers, given the complex social, economic and political factors at play; this challenge similarly applies to the other societal collapse proxies but is particularly noted in the migration studies. We observe from these studies that a food insecurity threshold for *natural mortality* is well established but thresholds for *conflict mortality* and *emigration* are not. Indeed, distinguishing causal drivers within datasets and defining quantitative thresholds for these determinants remains a "grand challenge".[94]



#	Proxy Examined	Typology Classification	Evidence Point Reference	Spatial Scale	Spatial Granularity	Temporal Period	Temporal Granularity
N1	Natural mortality	Model (SDM)	Meadows et al, 2004	Global	Global	1900-2004, fwd sim.	Yearly
N2		Statistical analysis	Agboola, 2017	Global (114 countries)	National	1995-2009	Yearly
N3			Lee et al, 2016	Global (95 countries)	National	2001-2011	Yearly
N4		Interview/survey	Campbell et al, 2009	Indonesia	Sub-national	2000-2003	Period
N5			Fledderjohann et al, 2016	India	Sub-national	2002-2008	Period
N6			Gundersen et al, 2018	Canada	Sub-national	2005, 2007-2010	Period
N7			Nandy et al, 2016	Nigeria, Ethiopia	Sub-national	2000-2013	Yearly
N8			Sati & Vangchhia, 2016	India	Sub-national	2014	Period
N9			Walker et al, 2019	USA	National	2003-10	Period
C1	Conflict mortality	Model (ABM)	Natalini et al, 2017	Global (213 countries)	National	2005-2013, fwd sim.	Monthly, yearly
C2		Statistical analysis	Arezki & Brueckner, 2014	Global (60 countries)	National	1970-2007	Yearly
C3			Bellemare, 2014	Global	Global	1990-2011	Monthly
C4			Berazneva & Lee, 2013	Africa (14 countries)	National	2007-2008	Period
C5			Caruso et al, 2016	Indonesia	Sub-national	1993-2003	Yearly
C6			Hendrix & Haggard, 2015	Africa, Asia (49 countries, 55 cities)	Sub-national	1961-2010	Yearly
C7			Jones et al, 2017	Africa	National	1991-2011	Monthly
C8			Koren & Bagozzi, 2016	Global	0.5×0.5 deg.	1991-2008	Yearly
C9			Maystadt et al, 2014	Global (arab world vs rest of world)	National	1960-2010	Yearly
C10			Natalini et al, 2015	Africa, Asia, MiddleEast (25 countries)	National	2005-2011	Monthly, yearly
C11			Newman, 2018	Global (79 countries)	National	2005-2015	Yearly
C12			Pinstrup-Andersen & Shimokawa, 2008	Global (146 countries)	National	1980-2005	Yearly
C13			Raleigh et al, 2015	Africa (24 countries, 113 markets)	Sub-national	1997-2010	Monthly
C14			Smith, 2014	Africa (40 countries)	National	1990-2012	Monthly
C15			van Weezel, 2016	Africa (45 countries)	National	1990-2011	Monthly
C16			Weinberg & Bakker, 2015	Global (71 countries)	National	1972-2007	Yearly
C17			Wischnath & Buhaug, 2014	India	Sub-national	1980-2011	Yearly
C18		Interview/survey	Legwegoh et al, 2015	Cameroon	Sub-national	2008	Period
C19		Data-led/scenario	Lunt et al, 2016	Global	National	2016, projection	Period
C20			Sternberg, 2012	Global (Egypt, China)	National	2010-2011	Period
E1	Emigration	Model (ABM)	Smith, 2014	Tanzania	Sub-national	2012, fwd sim.	Period, yearly
E2		Interview/survey	Afifi et al, 2014	Tanzania	Sub-national	2012	Period
E3			Etzold et al, 2014	Bangladesh	Sub-national	2011	Period
E4			Jacobson et al, 2019	Cambodia	Sub-national	2016	Period
E5			Milan & Ho 2014	Peru	Sub-national	2011	Period
E6			Milan & Ruano, 2014	Guatemala	Sub-national	2011	Period
E7			Murali & Afifi, 2014	India	Sub-national	2011	Period
E8			Panter-Brick & Eggerman, 1997	Nepal	Sub-national	1996 (assumed)	Period
E9			Rademacher-Schulz et al, 2014	Northern Ghana	Sub-national	2011	Period
E10			Warner & Afifi, 2014	Global (S/SEAsia, SSAfrica, LatAm)	Sub-national	2011-2012	Period
E11		Data-led/scenario	Doos, 1994	Global	Global, national	1990-94, projection	Yearly
E12			Puma et al, 2018	Global	National	2011-2018	Yearly

Number of evidence points
16
14
12
10
8
6
4
2
0
ABM
ABM
SDM
Natural Mortality
Conflict Mortality
Emigration
Legend
🟥 Complex systems model
🟦 Statistical analysis of dataset
🟩 Collection / analysis of survey data
🟨 Data-led case study / scenario
ABM = agent based model
SDM = system dynamics model

Fig. 4: Summary and custom colour-coded typology of the new empirical evidence base of climate change, food insecurity and societal collapse in contemporary society. A full reference list is contained in Supplementary Information (E.)

Each data-driven method offers different advantages. The complex systems models each describe "chunks" of the system at different scale and granularity. The models provide mathematical definition, are calibrated to real-world data and enable quantitative simulation of key relationships in the system. The statistical analyses quantitatively examine relationships between a dependent variable and one or more independent variables within the system, which can be used as a mathematical basis for extending modelling capabilities. The collection/

analysis of interview/survey data provides insight into qualitative aspects of human perspective and decision-making that quantitative data sets cannot provide directly. The data-led case study/scenarios combine quantitative data with qualitative expert interpretation to better understand global trends and forecasts. These latter two methods can also be used to inform the development of modelling capabilities, the scenarios analysed by such models and their application in decision-making processes. Collectively, these different data-driven methods can yield useful insights into the nuances of relationships in the system of interest.

4.2. Causal loop diagram of the climate change, food insecurity and societal collapse in contemporary society at global scale and national granularity

The main result of this chapter is the CLD (the f-CLD from Section 3, Step V), presented in Fig. 5. It structures the relationships between climate change, food insecurity and societal collapse as described in our new empirical evidence base (presented in Fig. 4 and discussed in Section 4.1). We discuss three key aspects of the CLD, namely insights related to the spread of empirical evidence, the qualitative complex system depicted, and quantitative complex system modelling, below, alongside consideration of well-established benefits and limitations of CLDs.

Our CLD is presented in a novel format that documents the spread of our empirical evidence base. We use line thickness and colour, respectively, to depict the density and type of the data-driven methods used by the empirical evidence points to analyse a given link between two variables.

Doing this aids comprehension of where existing work has been focused with respect to the climate change, food insecurity and societal collapse causal pathway. It may also help with the identification of gaps in existing analyses. For example, we can see that the link between *food insecurity* and *conflict* has been investigated mostly by evidence points using statistical analyses (blue), whereas the links between *food insecurity* and *migration*, and *food insecurity* and *natural mortality*, have been investigated mostly by evidence points using interviews/surveys (green). This hints that it may be useful to investigate the former using quantitative statistics, and the latter using qualitative interviews/surveys, to gain further insights offered by the different data-driven methods as described in Section 4.1.

It is important to recognise that our CLD may show negligible density for important links or even be missing important variables and/or links, either because they have not yet been studied or because our key word search failed to identify evidence points that have studied them. For example, our study focused on the climate change, food insecurity and societal collapse causal pathway, so the density of our empirical evidence is concentrated along links central to this pathway, whereas the links between peripheral variables in the system, such as between *fertility* and *births*, show a lower density of empirical evidence. Similarly, our use of the *population loss* set of societal collapse proxies means that the evidence base details *natural mortality, conflict mortality* and *emigration*, whereas the *institutional breakdown* set are not detailed. In considering this issue, our methodology attempted to maximise the rigour and transparency of our study by documenting the spread of our empirical evidence base to help make the reader aware of exactly how much and what type of evidence was supporting the CLD presented here.

Further, we can see that while empirical studies have linked *climate change* via *food insecurity* to our societal collapse proxies of *natural mortality, conflict mortality* and *emigration*, we found no empirical studies linking these proxies to the explicit term of *societal collapse*. This was expected given the motivation of this study (Section 1) and is due to the fact that there are no contemporary events of societal collapse, under the same definition as those in the historical studies pre-dating contemporary society, that enable these links to be empirically studied.[95]

Having considered the spread of empirical evidence, we now consider the complex system documented. A key benefit of CLDs is that they simply present a myriad of information in a single diagram; in doing so, CLDs enable comprehension of the structure and behaviour of complex systems, including feedbacks, intervention points and far-reaching interdependencies.[96] Our CLD visually depicts a system of 39 variables, 105 links and 32,000 feedback loops,[97] integrating information from different fields including climate science, food security, conflict, migration and health research.

Walking through the CLD at a high-level, we can see how *population* growth and *lifestyle emissions*, influenced by *institutional/demographic factors* (e.g. emission reduction incentives), combine to directly drive *climate change*. Similarly, they indirectly drive *climate change* via *consumer demand* on *food production*, which produces emissions directly (e.g. ruminant livestock) and indirectly via *industrial capital/output* (e.g. processing factories). The *environmental risk factors* (e.g. extreme weather events) of *climate change* may cause losses of *food production* either directly (e.g. plant disease) or indirectly via *agricultural input availability* (e.g. loss of water source for irrigation). A country's *food availability* is influenced by domestic *food production* and international *food trade. Food accessibility* is influenced by its *food price*, which responds to domestic (e.g. *cost of food production and distribution*) and international (e.g. *international food price*) markets, and *institutional/demographic factors* (e.g. food subsidies). *Food utilisation* is influenced by *infrastructure/services* (e.g. education) and *institutional/demographic factors* (e.g. cultural traditions). *Food insecurity* is underpinned by these three pillars of *food availability, food accessibility* and *food utilisation*. For a given country, *food insecurity* can drive *natural mortality* (i.e. *starvation*), *conflict* and *migration*, contributing to *population loss*, as well as economic shocks and socio-political instability, contributing to *institutional breakdown*, which exacerbates the risk of *societal collapse*.

Beyond a given country suffering increased *natural mortality*, famines (i.e. *food insecurity*) can place pressure on international humanitarian efforts (i.e. *institutional risk factors*). *Conflict* may occur domestically or internationally and can feedback to exacerbate *food insecurity* and institutional fragility (i.e. *institutional risk factors*). Potential mass *emigration* can increase pressure on *food availability, natural resources* and *infrastructure/services* in the destination nation, which can lead to socio-cultural tensions (i.e. *institutional risk factors*) that fuel *conflict. Food insecurity* can also directly contribute to *institutional risk factors* such as social unrest, political instability and economic inequality, which increase the risk of *societal collapse* due to *institutional breakdown*, that may also cascade internationally. While already fragile states are expected to be hit the worst directly, these insights reveal the indirect ramifications of climate change on our globalised society,[98] with serious consequences for humanity's "existential security".[99]

While some of these relationships may appear obvious, it is the act of bringing this information, which may otherwise be siloed and thus preventing consideration of the full story, together in one place that is of value.[100] In doing so, our CLD attempts to provide readers with the opportunity to explore the climate change, food insecurity and societal collapse causal pathway, consider worst-case scenarios that we want to avoid, develop transformative narratives of "where we want to go" and think about interventions that may help us attain this desired future.[101]

It is important to appreciate that CLDs are only as good as their information inputs; our CLD documents relationships based on information portrayed in our empirical evidence base as well as our interpretation of that information. As such, there exist challenges and limitations.[102] For instance, CLDs may mask variability of relationships in different contexts and locations, because they can only depict a single scale and granularity. The portrayal of explicit causality between variables in a CLD is a challenge as this can often work in both directions rather than one. CLDs can often become either too complicated or too simplified, which undermines their usefulness. In considering each of these issues, our original methodology attempted to maximise the rigour and transparency of our study by first documenting the information in each evidence point with an i-CLD and then consistently applying, and recording, the iterative process of reconciling the variables and links from each i-CLD to construct the f-CLD at the selected global scale and national granularity. In doing so, we sought to enable the reader to be aware of the nuances of the different scales and granularity of information underpinning our CLD, as well as our process of carefully reconciling causality, over 950 variables to 39 variables and 1150 links to 105 links to maximise the information conveyed while balancing readability.

It is also important to note that, due to their qualitative and static nature, CLDs do not enable us to comprehend the dynamics of the system, including nonlinear and emergent behaviour, non-intuitive quantitative results and time delays.[103] Complex systems models, although with their own challenges and limitations,[104] provide the opportunity to quantitatively analyse the dynamics of a system and gain insights into the potentially far-reaching impacts of our decisions.[105] However, complex systems models that explicitly examine societal collapse in contemporary society are underdeveloped. The World3 system dynamics model[106] — an evidence point in this study (refer to Supplementary Information D).[107] — is the eminent model of relevance, with only a limited number of studies building on it. World3 examines the potential for "overshoot-and-collapse" given population and industrial growth within the finite carrying capacity of the natural world system, implicitly accounting for climate change and explicitly accounting for food availability.

The information contained in our CLD and empirical evidence base may be useful in identifying and informing opportunities to improve these existing complex systems modelling capabilities for climate change, food insecurity and societal collapse scenarios. For example, our CLD highlights important factors at global scale and national granularity that World3 does not incorporate because it is defined at global scale and granularity.[108] World3 does not distinguish heterogenous characteristics of nations, such as distribution of *population* or geographic endowment of *natural resources*. It also does not account for international interactions, such as *food trade, conflict* and *migration*. Relatedly, World3 evaluates *societal collapse* only by *natural mortality* (defined by *food availability*, age and pollution) and does not include the other two *population loss* secondary determinants, as noted in the previous sentence, nor the three *institutional breakdown* secondary determinants. While our empirical evidence base may provide useful direction to datasets, it is important to note that quantitatively defining these relationships, particularly thresholds as discussed in Section 4.1, remains a key challenge of developing complex systems models. Nonetheless, given that individuals associate with national identity, business governance and policy-making are concentrated at national level, and international interactions underpin the functioning of contemporary society it could be valuable to model societal collapse risk profiles of different nations to inform the prioritisation and development of intervention strategies.

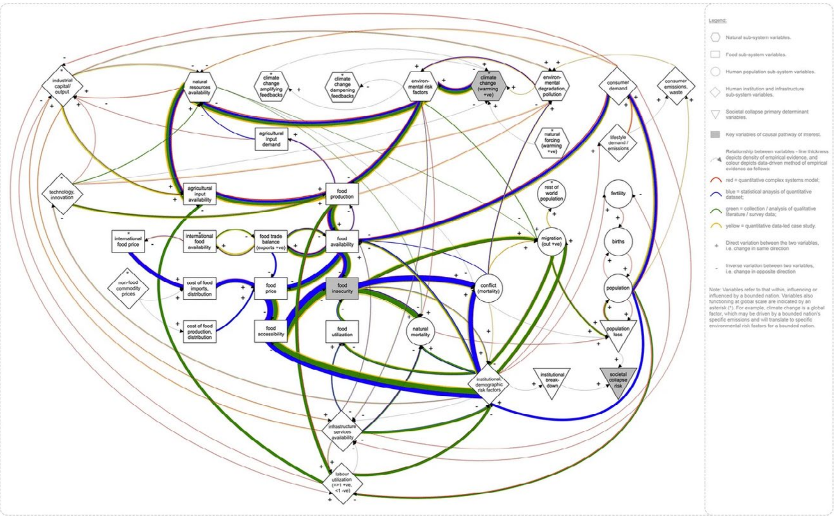

* industrial capital/ output
natural resources availability
* climate change amplifying feedbacks
* climate change dampening feedbacks
environ-mental risk factors
* climate change (warming +ve)
environ-mental degradation, pollution
consumer demand
consumer emissions, waste
agricultural input demand
* natural forcing (warming +ve)
lifestyle demand / emissions
technology, innovation
agricultural input availability
food production
* rest of world population
fertility
* international food price
* international food availability
food trade balance (exports +ve)
food availability
migration (out +ve)
births
* non-food commodity prices
cost of food imports, distribution
food price
food insecurity
conflict (mortality)
population
cost of food production, distribution
food accessibility
food utilization
natural mortality
population loss
institutional, demographic risk factors
institutional break-down
societal collapse risk
infrastructure services availability
labour utilization (=>1 +ve, <1 -ve)
Legend:
Natural sub-system variables.
Food sub-system variables.
Human population sub-system variables.
Human institution and infrastructure sub-system variables.
Societal collapse primary determinant variables.
Key variables of causal pathway of interest.
Relationship between variables - line thickness depicts density of empirical evidence, and colour depicts data-driven method of empirical evidence as follows:
red = quantitative complex systems model;
blue = statistical anaysis of quantitative dataset;
green = collection / analysis of qualitative literature / survey data;
yellow = quantitative data-led case study.
+ Direct variation between the two variables, i.e. change in same direction
- Inverse variation between two variables, i.e. change in opposite direction
Note: Variables refer to that within, influencing or influenced by a bounded nation. Variables also functioning at global scale are indicated by an asterisk (*). For example, climate change is a global factor, which may be driven by a bounded nation's specific emissions and will translate to specific environmental risk factors for a bounded nation.

Fig. 5: Causal loop diagram of climate change, food insecurity and societal collapse in contemporary society at global scale and national granularity. Variables are depicted as nodes in five different shapes, indicating different sub-systems. Links between variables are depicted as arrowed lines, indicating the direction of the relationship. Each link has a positive (+) or negative (-) notation, indicating that the two variables change in the same direction or opposite direction, respectively. The density and type of data-driven method of the empirical evidence base, from which the causal loop diagram was constructed, are depicted by line thickness and colour respectively.

5. Conclusions and Future Work

This chapter identified an empirical evidence base of climate change, food insecurity and societal collapse in contemporary society and structured the evidence base using a novel-format CLD defined at global scale and national granularity.

Two types of future work could extend from the results of this chapter. Identification of gaps in the spread of evidence across the CLD may guide future data-driven efforts to examine these causal relationships and define thresholds. The CLD and evidence base may be used to develop quantitative modelling capabilities, particularly by transforming the structure of World3 to account for heterogenous national characteristics and international interactions. Three types of future work could extend from the methodology and literature synthesis. The causal pathway examined in this chapter could be further detailed by re-applying the methodology using the *institutional breakdown* set of societal collapse proxies instead of the *population loss* set. The methodology, using either set of societal collapse proxies, could be applied to detail other causal pathways between climate change and societal collapse. The methodology, excluding the contemporary time-period limitation, could be applied to document the information in the historical studies identified in the literature review. Similarly, the methodology could be applied to construct CLDs at different scales and granularities.

It is hoped that this chapter has contributed to developing our understanding of the causal pathways through which climate change poses an existential risk to humanity and facilitates opportunities for future work.

Notes and References

[1] Ripple, W. J., C. Wolf, T. M. Newsome et al. 'World scientists' warning of a climate emergency', *Bioscience, 70* (2019): 8–12. https://doi.org/10.1093/biosci/biz088

[2] Raftery, A. E., A. Zimmer, D. M. W. Frierson et al. 'Less than 2 °C warming by 2100 unlikely', *Nat Clim Chang, 7* (2017): 637–41. https://doi.org/10.1038/nclimate3352

[3] Natali, S. M. 'Large loss of CO2 in winter observed across the northern permafrost region', *Nat Clim Chang*, 9 (2019): 852–57. https://doi.org/10.1038/s41558-019-0592-8

[4] Walker, X. J., J. L. Baltzer, S. G. Cumming et al. 'Increasing wildfires threaten historic carbon sink of boreal forest soils', *Nature, 572* (2019): 520–23. https://doi.org/10.1038/s41586-019-1474-y

[5] Steffen, W., J. Rockström, K. Richardson et al. 'Trajectories of the earth system in the Anthropocene', *Proc Natl Acad Sci USA, 115* (2018): 8252–59. https://doi.org/10.1073/pnas.1810141115

[6] Schneider, T., C. M. Kaul and K. G. Pressel. 'Possible climate transitions from breakup of stratocumulus decks under greenhouse warming', *Nat Geosci, 12* (2019): 163–67. https://doi.org/10.1038/s41561-019-0310-1

[7] The Center for Climate & Security. 'A security threat assessment of global climate change: how likely warming scenarios indicate a catastrophic security future', *The National Security, Military and Intelligence Panel on Climate Change*. The Center for Climate and Security (2020).

[8] Xu, C., T. A. Kohler, T. M. Lenton et al. 'Future of the human climate niche', *Proc Natl Acad Sci USA, 117* (2020): 11350–55. https://doi.org/10.1073/pnas.1910114117

[9] Ord, T. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing (2020).

[10] Kunreuther, H., G. Heal, M. Allen et al. 'Risk management and climate change', *Nat Clim Chang, 2* (2013): 447–50. https://doi.org/10.1038/NCLIMATE1740

11 Shepherd, T. G., E. Boyd, R. A. Calel et al. 'Storylines: An alternative approach to representing uncertainty in physical aspects of climate change', *Clim Chang, 151* (2018): 555–71. https://doi.org/10.1007/s10584-018-2317-9

12 Briggs, S., C. F. Kennel and D. G. Victor. 'Planetary vital signs', *Nat Clim Chang, 5* (2015): 969–70. https://doi.org/10.1038/nclimate2828

13 Wagner, G. and M. L. Weitzman. *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton University Press (2015).

14 Weber, E. U. 'Evidence-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet)', *Clim Chang, 77* (2006): 103–20. https://doi.org/10.1007/s10584-006-9060-3

15 Gowdy, J. 'Our hunter-gatherer future: Climate change, agriculture and uncivilization‘, *Futures, 115* (2020): 102488. https://doi.org/10.1016/j.futures.2019.102488

16 FAO, IFAD, UNICEF, et al. *The State of Food Security and Nutrition in the World 2020: Transforming Food Systems for Affordable Healthy Diets*. Rome (2020).

17 Rivington, M., R. Bailey, T. Benton et al. *Extreme Weather and Resilience of the Global Food System Synthesis Report* (2015).

18 IPCC. 'Climate change 2014: Synthesis report', *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (2014).

19 Butzer, K. W. 'Collapse, environment and society', *Proc Natl Acad Sci USA, 109* (2012): 3628–31. https://doi.org/10.1073/pnas.1114845109

20 Kemp, L. 'Climate endgame: understanding worst-case warming', in F. Riede, M. Evans and L. Harrington (eds.), *Learning From the Past Working Group Webinar. Future Earth's Knowledge Action Network on Emergent Risks and Extreme Events* (2020).

21 Randers, J. 'From limits to growth to sustainable development or SD (sustainable development) in a SD (system dynamics) perspective', *Syst Dyn Rev, 16* (2000): 213–24. https://doi.org/10.1002/1099-1727(200023)16:3<213::AID

22 Ord (2020).

23 Kemp, L. 'Are we on the road to civilization collapse?', *BBC Future* (2019).

24 Butzer, K. W. and G. H. Endfield. 'Critical perspectives on historical collapse', *Proc Natl Acad Sci USA, 109* (2012): 3628–31. https://doi.org/10.1073/pnas.1114772109

25 de Menocal, P. B. 'Cultural responses to climate change during the late holocene', *Science, (80-)*292 (2001): 667–73. https://doi.org/10.1126/science.1059827

26 Weiss, H. and R. S. Bradley. 'What drives societal collapse?', *Science, (80-)*291 (2001): 609–10.

27 Hodell, D., J. H. Curtis and M. Brenner. 'Possible role of climate in the collapse of classic Maya civilization', *Nature, 375* (1995): 391–94. https://doi.org/10.1038/375391a0

28 Haug, G. H., D. Günther, L. C. Peterson et al. 'Climate and the collapse of Maya civilization', *Science, (80-)*299 (2003): 1731–35. https://doi.org/10.1126/science.1080444

29 Medina-Elizalde, M. and E. J. Rohling. 'Collapse of classic Maya civilization related to modest reduction in precipitation', *Science, (80-)*335 (2012): 956–59. https://doi.org/10.1126/science.1216629

30 Weiss, H., M.-A. Courty, W. Wetterstrom et al. 'The genesis and collapse of third millennium north Mesopotamian civilization', *Science, (80-)*261 (1993): 995–1004. https://www.science.org/doi/10.1126/science.261.5124.995

31 Cullen, H. M., P. B. de Menocal, S. Hemming et al. 'Climate change and the collapse of the Akkadian empire: Evidence from the deep sea', *Geology, 28* (2000). https://doi.org/10.1130/0091-7613(2000)28<379:CCATCO>2.0.CO;2

32 Cookson, E., D. J. Hill and D. Lawrence. 'Impacts of long term climate change during the collapse of the Akkadian empire', *J Archaeol Sci, 106* (2019): 1–9. https://doi.org/10.1016/j.jas.2019.03.009

33 Cline, E. H. *1177 BC: The Year Civilization Collapsed*. Princeton University Press (2014).

34 Finné, M., K. Holmgren, C.-C. Shen et al. 'Late bronze age climate change and the destruction of the Mycenaean palace of Nestor at Pylos', *PLoS One, 12* (2017). https://doi.org/10.1371/journal.pone.0189447

35 Giosan, L., P. D. Clift, M. G. Macklin et al. 'Fluvial landscapes of the Harappan civilization', *Proc Natl Acad Sci USA, 109* (2012): E1688–E1694. https://doi.org/10.1073/pnas.1112743109

36 Buckley, B. M., K. J. Anchukaitis, D. Penny et al. 'Climate as a contributing factor in the demise of Angkor, Cambodia', *Proc Natl Acad Sci USA, 107* (2010): 6748–52. https://doi.org/10.1073/pnas.0910827107

37 Zhang, D. D., C. Jim, G.-S. Lin et al. 'Climatic change, wars and dynastic cycles in China over the last millennium', *Clim Chang, 76* (2006): 459–77. https://doi.org/10.1007/s10584-005-9024-z

38 Li, Z., Y. Chen, Y. Wang and L. Weihong. 'Drought promoted the disappearance of civilization along the ancient Silk Road', *Environ Earth Sci, 75* (1116) (2016). https://doi.org/10.1007/s12665-016-5925-6

39 Kintisch, E. 'Why did Greenland's Vikings disappear?', *Science, 354*(6313) (2016): 696–701.

40 Ortloff, C. R. and A. L. Kolata. 'Climate and collapse: Agro-ecological perspectives on the decline of the Tiwanaku state', *J Archaeol Sci, 20* (1993): 195–221. https://doi.org/10.1006/jasc.1993.1014

41 Ehrlich, P. and A. Ehrlich. 'Can a collapse of global civilization be avoided?', *Proc R Soc B, 280* (2013). https://doi.org/10.1098/rspb.2012.2845

42 Fagan, B. M. *The Great Warming: Climate Change and the Rise and Fall of Civilizations*. Bloomsbury Press (2008).

43 McMichael, A. J. *Climate Change and the Health of Nations: Famines, Fevers and the Fate of Populations*. Oxford University Press (2017).

44 Runciman, W. G. 'The origin of human social institutions', *Proc Br Acad, 110* (2001).

45 Ponting, C. *A New Green History of the World: The Environment and the Collapse of Great Civilizations*. Vintage (1991).

46 Wright, R. *A Short History of Progress*. House of Anansi Press Inc. (2004).

47 Diamond, J. *Collapse: How Societies Choose to Fail or Succeed*. Penguin Group (2005).

48 Malthus, T. R. *An Essay on the Principle of Population*. Cambridge University Press (1978).

49 Hardin, G. 'The tragedy of the commons', *Science, (80-)*162 (1968): 1243–48. https://doi.org/10.1126/science.162.3859.1243

50 Catton, W. R. *Overshoot: The Ecological Basis of Revolutionary Change*. Board of Trustees of the University of Illinois (1980).

51 Toynbee, A. J. *A Study of History* (*Vol I — IX*). Oxford University Press (1961); Kempt (2019).

52 Tainter, J. A. *The Collapse of Complex Societies*. Cambridge University Press (1988).

53 Acemoglu, D. and J. A. Robinson. *Why Nations Fail: The Origins of Power, Prosperity, and Poverty*. Crown Publishers (2012).

54 Johnson, S. A. J. *Why Did Ancient Societies Fail*? Routledge (2017).

55 Tainter (1988).

56 Murphy, D. J. and C. A. S. Hall. 'Year in review — EROI or energy return on (energy) invested', *Ann N Y Acad Sci, 1185* (2010): 102–18. https://doi.org/10.1111/j.1749-6632.2009.05282.x

57 Homer-Dixon, T. *The Upside of Down: Catastrophe, Creativity, and the Renewal of Civilization*. Random House of Canada (2006).

58 Diamond J (1997) Guns, germs and steel: a short history of everybody for the last 13,000 years. Vintage, London.

59 Turchin P (2006) War and peace and war: the rise and fall of empires. Penguin Group, New York.

60 Schwartz, G.M. and J. J. Nichols. *After Collapse: The Regeneration of Complex Societies*. University of Arizona Press (2010).

61 Gibbon, E. *The History of the Decline and Fall of the Roman Empire* (*Vol I — VI*). Strahan & Caddell (1789).

62 Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable*. Random House (2007).

63 Arbesman, S. 'The life-spans of empires', *Hist Methods A J Quant Interdiscip Hist, 44* (2011): 127–29. https://doi.org/10.1080/01615440.2011.577733

64 Van Valen, L. 'A new evolutionary law', *Evol Theory, 1* (1973): 1–30.

65 Bostrom, N. and M. M. Ćirković. *Global Catastrophic Risks*. Oxford University Press (2008).

66 Rees, M. *On the Future: Prospects for Humanity*. Princeton University Press (2018).

67 Ord (2020).

68 WEF. *The Global Risks Report 2020* (2020).

69 GCF. *Global Catastrophic Risks 2020*. Global Challenges Foundation (2020).

70 Ehrlich and Ehrlich (2013).

71 Häggström, O. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford University Press (2016).

72 Rivington et al. (2015).

73 Morton, O. *The Planet Remade: How Geoengineering Could Change the World*. Princeton University Press (2016).

74 Homer-Dixon, T., B. Walker, R. Biggs et al. 'Synchronous failure: The emerging causal architecture of global crisis', *Ecol Soc, 20*(6) (2015). https://doi.org/10.5751/ES-07681-200306

75 Buldyrev, S. V., R. Parshani, G. Paul et al. 'Catastrophic cascade of failures in interdependent networks', *Nature, 464* (2010): 1025–28. https://doi.org/10.1038/nature08932

76 Lynas, M. *Six Degrees: Our Future on a Hotter Planet*. Fourth Estate (2007).

77 Wallace-Wells, D. *The Uninhabitable Earth: A Story of the Future*. Crown Publishing Group (2019).

78 Gowdy (2020).

79 Mora, C., B. Dousset, I. R. Caldwell et al. ‘Global risk of deadly heat’, *Nat Clim Chang, 7* (2017): 501–6. https://doi.org/10.1038/nclimate3322

80 Hsiang, S. M., M. Burke and E. Miguel. ‘Quantifying the influence of climate on human conflict’, *Science (80-), 496* (2013). https://doi.org/10.1126/science.1235367

81 Hauer, M. E., E. Fussell, V. Mueller et al. ‘Sea-level rise and human migration‘, *Nat Rev Earth Environ, 1* (2020): 28–39. https://doi.org/10.1038/s43017-019-0002-9

82 Abel, G. J., M. Brottrager, J. C. Cuaresma and R. Muttarak. ‘Climate, conflict and forced migration’, *Glob Environ Chang, 54* (2019): 239–49. https://doi.org/10.1016/j.gloenvcha.2018.12.003

83 Burke, M., S. M. Hsiang and E. Miguel. ‘Global non-linear effect of temperature on economic production’, *Nature, 527* (2015): 235–39. https://doi.org/10.1038/nature15725

84 Sofuoğlu, E. and A. Ay. ‘The relationship between climate change and political instability: The case of MENA countries (1985:01–2016:12)’, *Environ Sci Pollut Res* (2020). https://doi.org/10.1007/s11356-020-07937-8

85 Adger, W. N., J. Barnett, K. Brown et al. ‘Cultural dimensions of climate change impacts and adaptation’, *Nat Clim Chang, 3* (2013): 112–17. https://doi.org/10.1038/nclimate1666

86 Theisen, O. M., N. P. Gledistch and M. Buhaug. ‘Is climate change a driver of armed conflict?’, *Clim Chang, 117* (2013): 613–25. https://doi.org/10.1007/s10584-012-0649-4

87 Gemenne, F., J. Barnett, W. N. Adger and G. D. Dabelko. ‘Climate and security: Evidence, emerging risks, and a new agenda‘, *Clim Chang, 123* (2014): 1–9. https://doi.org/10.1007/s10584-014-1074-7

88 Bostrom, N. ‘Existential risk prevention as global priority’, *Glob Policy, 4* (2013): 15–31. https://doi.org/10.1111/1758-5899.12002

89 Meadows, D. H., D. L. Meadows and J. Randers. *The Limits to Growth: The 30-Year Update*. Chelsea Green Publishing Company (2004).

90 e.g. IPCC (2014).

91 FAO. ‘Human energy requirements’, *Report of a Joint FAO/WHO/UNU Expert Consultation, Rome* (2004).

92 UCDP/PRIO. *Uppsala Conflict Data Program* (2019). https://ucdp.uu.se/

93 IOM GMDAC. *Migration Data Portal: The Bigger Picture* (2019). https://migrationdataportal.org

94 Kintigh, K. W., J. H. Altschul, M. C. Beaudry et al. ‘Grand challenges for archaeology’, *Proc Natl Acad Sci, 111* (2014): 879–80. https://doi.org/10.7183/0002-7316.79.1.5

95 Beard, S., T. Rowe and J. Fox. ‘An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards’, *Futures, 115* (2020): 102469. https://doi.org/10.1016/j.futures.2019.102469

96 Sterman, J. D. ‘Communicating climate change risks in a skeptical world’, *Clim Chang, 108* (2011): 811. https://doi.org/10.1007/s10584-011-0189-3

97 Calculated using Vensim ® PLE software, Ventana Systems Inc.

98 Kemp (2020).

99 Sears, N. A. 'Existential security: Towards a security framework for the survival of humanity', *Glob Policy, 11* (2020): 255–66. https://doi.org/10.1111/1758-5899.12800

100 Sterman (2011).

101 Hinkel, J., D. Mangalagiu, A. Bisaro and J. D. Tàbara. 'Transformative narratives for climate action', *Clim Chang, 160* (2020): 495–506. https://doi.org/10.1007/s10584-020-02761-y

102 Cernev, T. and R. Fenner. 'The importance of achieving foundational sustainable development goals in reducing global risk', *Futures, 115* (2020): 102492. https://doi.org/10.1016/j.futures.2019.102492

103 Sterman (2011).

104 Sterman, J. D. 'All models are wrong: Reflections on becoming a systems scientist', *Syst Dyn Rev, 18* (2002): 501–31. https://doi.org/10.1002/sdr.261

105 Sterman, J. D. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. McGraw-Hill Higher Education (2000).

106 Meadows et al. (2004).

107 E.g Ansell, T. and S. Cayzer. 'Limits to growth redux: A system dynamics model for assessing energy and climate change constraints to global growth', *Energy Policy* (2018): 514–25. https://doi.org/10.1016/j.enpol.2018.05.053

108 Meadows et al. (2004).

14. Existential Change: Lesson from Climate Change for Existential Risk

SJ Beard and Luke Kemp

Highlights:

- In this short chapter the authors draw on several research strands and papers within CSER to offer a theoretical reflection on how to think about catastrophic climate change and what Existential Risk Studies can learn from climate change research.
- This is intended to build on the previous chapter, in which Catherine Richards, Richard Lupton, and Julian Allwood provide an empirical assessment of one highly concerning risk cascade involving climate change and highlight its potential contribution to global catastrophic and existential risk.
- Climate change is one of the most empirically well-studied risks and has deep links to pre-existing bodies of literature, such as disaster risk management, environmental studies, and food security.
- Drawing on these studies and more, the chapter reflects on how to frame research questions in existential risk, what causes catastrophic climate change to be neglected by climate and existential risk researchers alike, and how to incorporate assessments of response risk and co-benefits into thinking about catastrophic climate change.

https://doi.org/10.11647/OBP.0360.14

This short chapter brings together a number of important ideas and draws readers attention to other extant bodies of literature. The relative value of co-benefits approaches is discussed in other chapters in this volume, including Chapter 4, in more detail. The dangers of response risks are further discussed in Chapter 2.

1. Asking the Wrong Questions for the Right Reasons

Within Existential Risk Studies it is common to hear people ask the question "is climate change an existential risk?", and many who ask this question answer negatively, arguing that as a result climate change is not an important topic of research within the field. However, whether it is answered affirmatively or not, this question is misguided. There are three reasons for thinking this. Firstly, it makes little sense on a probabilistic level; whether something will be a threat to our collective existence is not a binary matter, it is a question of likelihood. However, many researchers within Existential Risk Studies mistakenly conflict existential risk with events that could be existential catastrophes. Secondly, climate change is not a single uniform process that will affect everyone in the same way; it is a set of diffuse impacts to different exposed populations, interacting with different vulnerabilities and exposures, and activating different risk cascades. As Richards et al. show, it will inevitably interact with a host of other threats (not only food security and societal collapse, but even factors such as the explosivity of volcanic eruptions or the emergence of zoonotic pathogens),[1] and these can interact with one another to create reinforcing feedback loops or "global systems death spirals".[2] Finally, "existential risk" is too vague and arbitrary a concept for the question to ever be answered. All the definitions of existential risk that have received the greatest public attention thus far, such as Toby Ord's, focused not in terms of an impact on humanity at any point in time but rather in terms of "the loss of long-term future value";[3] either referring to the author(s) particular vision of a high-tech intergalactic utopia, or a fuzzy undefined idea of "our potential".[4]

Other authors have practised attribution substitution and sought to answer an easier question such as "will the direct impacts of climate change make the Earth uninhabitable?" as a proxy for existential risk,[5] or suggested agricultural impossibility as a proxy for civilisational collapse at a given level of temperature rise.[6] These are certainly more tractable questions, but they are also entirely different questions, and there is a danger in thinking that answering them is sufficient to assess the overall level of climate risk.

We are better off reverting back to the common-sense definition of existential risk as the risk to the existence of a given object, and specifying whether the object under threat is humanity as a whole (extinction risk), global industrial society (collapse risk), or something else entirely. We should be thinking of an overall level of risk emergent from a particular socio-ecological system, and how much climate change influences this level.[7] And the question we should be asking about this risk is what contribution, under certain scenarios, climate change will make, bearing in mind that it will almost certainly be operating in tandem with many other drivers of risk.

Considering this revised question can also help to rectify a recurring problem in the climate risk literature: using mean global temperature rise as the sole threat indicator. Authors and activists alike have frequently made a direct link between the level of warming and the likelihood of global catastrophe, with 4–6 °C being most frequently used as this terrible threshold.[8] However, global surface temperature is only one of the climate change induced factors we need to worry about. 3 °C of warming above pre-industrial levels could be entirely manageable if it occurs in a world of adaptive technologies, high levels of multilateral cooperation, wealth equality, trust in institutions, and the safe management of other planetary boundaries. It could also be catastrophic in a world where other planetary boundaries are transgressed, the international order is riven with conflict, lethal autonomous weapons are in mass production, and societies are scarred by inequality, low trust, and polarisation. Understanding the contribution of climate change to Global Catastrophic Risk requires a more sophisticated approach which looks beyond the direct impacts of a given level of warming to think through fully formed climate scenarios. We believe that, when conceived of in this way, the risks associated with climate change are more appreciable and it is far harder to argue that understanding them is unimportant; however, even if others disagree with this assessment, we still maintain that this is the right way to think about the problem.

2. Catastrophic Neglect

Given how poorly questions about catastrophic climate change are often framed, it is hardly surprising that it has been a highly neglected subject of study, not only among existential risk researchers but also among climate change researchers. Even at the basic level of temperature rise scenarios, we give far more attention to studying the impacts of lower-end warming rather than high-end warming. Text-mining of IPCC reports shows that mentions of 3 °C and above is underrepresented relative to its likelihood (and impact),[9] a finding that has been verified by both literature sampling and the reports of popular authors trying to summarise the climate risk science.[10] If anything, this trend appears to have worsened over time with subsequent IPCC reports.1[11] The use of complex risk assessments to study climate scenarios has also been neglected: looking at compound hazards is already rare,[12] let alone considering risk cascades and integrated climate catastrophe assessments. Yet catastrophic climate change remains high on the public and political agenda, creating both a perception that this is a risk receiving far more attention than it is, and also an intellectual vacuum that is easily filled by poor quality research, ranging from speculative doom-mongering[13] to overly simplistic neoclassical economic models.[14]

There are four key reasons for this oversight of extreme global climate risk. First is international climate policy. The 2015 Paris Climate Agreement on Climate Change has channelled scientific attention toward the agreement's goal of limiting warming to 2 °C above pre-industrial levels and pursuing efforts to stabilise it below 1.5 °C, as these are now the publicly stated goals of climate negotiations (even if they are highly unlikely to actually be realised). Second, analysis of high-end warming scenarios and complex risk assessments are simply harder to do. The higher the warming gets, the more difficult it becomes to study, as these scenarios are more displaced for the current climatic niche. Moreover, complex climate risk assessments involving multiple factors are far more challenging than a hazard-centric analysis focusing on only the direct impacts of mean global temperature rise. Third, climate scholarship has had a strong incentive to "err on the side of least drama".[15] Climate change has long been the target of fossil-fuel industry campaigns to sow doubt, not just on attempts to assess climate change's catastrophic potential but even the fundamental science, and this creates incentives for conservative science that builds consensus and does not risk exploring divergent hypotheses.[16] Finally, many fear that discussing extreme risk could cause people to dwell too much on worst case scenarios, breeding fatalism and paralysis. However, this concern is misplaced; meta-analyses over hopeful vs. fearful messaging are mixed,[17] and in any case this is a false dichotomy. One of the most referenced pieces for those concerned about the paralytic effect of fear does show that hopeful messaging is more poignant than fear but also that "worry" is even more effective than hope.[18] The difference between worry and fear is one of degrees; the latter could even dissipate into the former over time. Furthermore, research should not be a PR exercise aimed to sway the public, in open democracies we have a duty to do honest risk assessments combined with clear recommendations for what can be done.[19]

Of course, these factors are only compounded by the consensus procedures of the IPCC, which seeks to synthesise scientific evidence for political purposes but is still often held up as a neutral arbiter of climate science. While useful, these procedures tend to produce lowest common-denominator outcome, which is precisely what is not needed when exploring extreme risks.[20] This is an important point of reflection for any future efforts to build similar bodies aimed at bringing scientific research to bear on the governance of other global risks.

3. The Risks and Rewards of Responding

Climate change is inherently tractable and we already have the technologies we need to stop creating it, albeit without the institutions to fairly distribute them with a sufficient level of urgency. However, responding to risks like climate change can incur risks of its own. Indeed, the IPCC, in its risk concept notes to the sixth assessment report, does not just discuss the usual three determinants of risk, hazard, vulnerability, and exposure, but also identifies "response risks".[21] Others have suggested that response should be added to the classic list of determinants.[22] In some cases, responses may be far worse than the initial perceived risk, that is, they are iatrogenic: the treatment is worse than the disease.

Existential risk is especially prone to response risks due to its scale, severity, and often speculative nature. For instance, at the extreme a speculative fear of dispersed weapons of mass destruction could justify a mass surveillance state.[23] In general, there is always the potential for concerns over global risk to justify a Stomp Reflex — the abuse of emergency powers which inappropriately empower those atop a hierarchy and shield them from scrutiny. [24] This is also true for climate change

Reacting to climate change could lead to emergency responses, such as stratospheric aerosol injection (SAI), in an attempt to manipulate the quantity of solar radiation hitting the earth and thus counter some of the impacts of climate change. Existing data on the direct impacts of SAI and its contribution to systemic risk or triggering other hazards is sparse. Preliminary analysis suggests that the greatest problem is the latent risks of "termination shock". If a calamity such as a nuclear war deactivates the system for a prolonged time, then this could significantly accelerate warming. Hence SAI shifts the risk distribution by likely lowering the level of risk in an average scenario but fattening the tail or "worst-case" scenarios depending on how SAI is deployed, to what degree it is used, and what geopolitical and ecological world it is dispersed into.[25] On the other hand, there are also frequently neglected co-benefits of climate mitigation policies, such as the public health benefits of eliminating coal smoke and other pollutants from our air.[26]

Such problems of response risk are perhaps the most neglected. Yet they are precisely what the study of existential risk needs to grapple with. This could include by using robust decision-making procedures, such as the minimax principle, to aid in selecting policy options under uncertainty or using deliberative democratic processes to combine diverse perspectives and co-create effective policy responses.

Notes and References

[1] IPCC. *Climate Change 2022: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press (2022).

[2] Beard, SJ, Lauren Holt, Asaf Tzachor, Luke Kemp, Shahar Avin, Phil Torres and Haydn Belfield. 'Assessing climate change's contribution to Global Catastrophic Risk', *Futures* , *127* (2021). https://doi.org/10.1016/j.futures.2020.102673

[3] Ord, T. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Publishing (2020); Cotton-Barratt, O. and T. Ord. *Existential Risk and Existential Hope: Definitions* (2015); Bostrom, N. 'Existential risk prevention as global priority', *Glob. Policy, 4* (2013): 15–31; Bostrom, N. 'Existential risks: Analysing human extinction scenarios and related hazards', *J. Evol. Technol., 9* (2002): 1–36.

[4] Cremer, C. Z. and L. Kemp. 'Democratising risk: In search of a methodology to study existential risk', *SSRN Electron. J.* (2021): 1–35.

[5] Ord (2020).

[6] MacAskill (2022).

[7] For the definition of other key terms, see Kemp, L. et al. 'Climate endgame: Exploring catastrophic climate change scenarios', *Proceedings of the National Academy of Sciences,. 119* (2022). https://doi.org/10.1073/pnas.2108146119

[8] Lynas, M. *Our Final Warning: Six Degrees of Climate Emergency*. Harper Collins (2020); Lynas, M. *Six Degrees: Our Future on a Hotter Planet*. National Geographic (2007); Wagner, G. and M. L. Weitzman. *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton University Press (2015); MacAskill, W. *What We Owe the Future*. Oneworld Publications (2022); Ord (2020).

[9] Jehn, F. U., M. Schneider, J. R. Wang, L. Kemp and L. Breuer. 'Betting on the best case: Higher end warming is underrepresented in research', *Environmental Research Letters, . 16* (2021): 084036. https://doi.org/10.1088/1748-9326/ac13ef

[10] Wallace-Wells, D. *The Uninhabitable Earth*. Crown Publishing Group (2019); Lynas (2020).

[11] Jehn, F. U. et al. 'Focus of the IPCC assessment reports has shifted to lower temperatures', *Earth's Future, 10* (2022). https://doi.org/10.1029/2022ef002876

[12] Matthews, T., R. L. Wilby and C. Murphy. 'An emerging tropical cyclone — Deadly heat compound hazard', *Nature Climate Change, 9* (2019): 602–6. https://doi.org/10.1038/s41558-019-0525-6

[13] Bendell, J. *Deep Adaptation: A Map for Navigating Climate Tragedy* (2018).

[14] Nordhaus, William D. 'The economics of tail events with an application to climate change', *Review of Environmental Economics and Policy* (2011). https://doi.org/10.1093/reep/rer004

[15] Brysse, K., N. Oreskes, J. O'Reilly and M. Oppenheimer. 'Climate change prediction: Erring on the side of least drama?', *Global Environmental Change, 23* (2013): 327–37. https://doi.org/10.1016/j.gloenvcha.2012.10.008

[16] Oreskes, N. and E. M. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues From Tobacco Smoke to Global Warming*. Bloomsbury Press (2010).

[17] Peters, G.-J. Y., R. A. C. Ruiter and G. Kok. 'Threatening communication: A critical re-analysis and a revised meta-analytic test of fear appeal theory', *Health Psychology Review, 7* (2013): 8–31. https://doi.org/10.1080/17437199.2012.703527; Tannenbaum, M. B. et al. 'Appealing to fear: A meta-analysis of fear appeal effectiveness and theories', *Psychological Bulletin, 141* (2015): 1178–1204. https://doi.org/10.1037/a0039729.supp

[18] Smith, N. and A. Leiserowitz. 'The role of emotion in global warming policy support and opposition', *Risk Analysis, 34* (2014): 937–48. https://doi.org/10.1111/risa.12140

[19] Kemp, L. et al. 'Reply to Bhowmik et al: Democratic climate action and studying

extreme climate risks are not in tension', *Proc. Natl. Acad. Sci.* (2022).

20 Kemp, L. 'Framework for the future? Exploring the possibility of majority voting in the climate negotiations', *Int. Environ. Agreements Polit. Law Econ., 16* (2016): 757–79.

21 Reisinger, Andy, Mark Howden, Carolina Vera et al. *The Concept of Risk in the IPCC Sixth Assessment Report: A Summary of Cross-Working Group Discussions* (2020).

22 Simpson, N. P. et al. 'A framework for complex climate change risk assessment', *One Earth, 4* (2021): 489–501. https://doi.org/10.1016/j.oneear.2021.03.005

23 A proposal that has been seriously mooted by some, e.g. Bostrom, Nick. 'The vulnerable world hypothesis', *Global Policy*, *10*(4) (2019): 455–76.

24 Kemp, L. 'The Stomp Reflex: When governments abuse emergency powers', *BBC Future* (2021).

25 Tang, A. and L. Kemp. 'A fate worse than warming? Stratospheric aerosol injection and catastrophic risk', *Front. Clim. Sci.* (2021): 1–17. https://doi.org/10.3389/fclim.2021.720312

26 West, Jason et al. 'Co-benefits of global greenhouse gas mitigation for future air quality and human health', *Nature Climate Change, 3*(10) (2013): 885–89. https://doi.org/10.1088/1748-9326/aa8f76

15. A Fate Worse Than Warming? Stratospheric Aerosol Injection and Catastrophic Risk

Aaron Tang and Luke Kemp

Highlights:

- This chapter considers the potential impact on Global Catastrophic Risk of injecting particles into the atmosphere to reflect sunlight, stratospheric aerosol injection (SAI). This both represents a potential technological solution to the threat of climate change and a contributor to GCR in its own right.
- Analyses of potential high impact outcomes from SAI are lacking in contemporary research. This chapter helps resolve this gap by investigating four aspects of SAI's potential contributions to catastrophic risk: 1) acting as a direct catastrophic risk (through ecological blowback); 2) interacting with other globally catastrophic hazards like nuclear war; 3) exacerbating other risks that cascade and amplify across different systems; and 4) acting as a latent risk that is dormant but can later be triggered.
- It finds that: the potential for major unforeseen environmental consequences seems highly unlikely but is ultimately unknown; SAI plausibly interacts with other catastrophic calamities, most notably nuclear war or an extreme space weather event; SAI could contribute to systemic risk by introducing stressors into critical systems such as agriculture

https://doi.org/10.11647/OBP.0360.15

but this is highly understudied; and SAI deployment more tightly couples different ecological, economic, and political systems, creating a precarious condition of latent risk that is the largest cause for concern.

- Across all these dimensions, the specific SAI deployment and associated governance, is critical. A well-coordinated use of a small amount of SAI could incur negligible risks, but this is an optimistic scenario. Conversely, larger use of SAI in an uncoordinated manner poses many potential dangers. We cannot equivocally determine whether SAI will be worse than warming. For now, a heavy reliance on SAI seems an imprudent policy response.

This chapter provides a detailed case study of how and why interventions into climatic change require rigorous analysis. Without investigation into the possible harms precipitated by technological intervention, there is a very real risk that unforeseen consequences could be dramatic. The lessons drawn from this case study are likely instructive for other areas of GCR research, and other examples of generalisable case study research in the field can be found in Chapters 13 and 16.

1. Introduction: Hothouse Earth or Shithouse Earth?

Could the risks of large-scale solar geoengineering be worse than the dangers posed by climate change? Many concerns have been expressed over geoengineering the Earth's climate. These tend to centre on solar radiation management (SRM) methods, particularly stratospheric aerosol injection (SAI). These range from fears over negative, unintended effects on ecology, political conflict, mitigation deterrence to ethical objections. Given the breadth of objections, it is quite clear that SAI would be iatrogenic in some way. Like some medical interventions, SAI may have adverse side-effects and complications. The question is whether it could be worse than the problem it is seeking to remedy: climate change.

There is a wealth of information on the different risks posed by climate change (although notably little on high-end warming scenarios), yet few attempts to compare this to the potential damages of SAI. This is unsurprising since there have been limited attempts to systematically analyse the myriad of threats posed by SAI.

We address this gap by analysing the severe downside risks of SAI. We do not directly compare the risks posed by SAI and climate change in this chapter. Rather, we provide an analytical foundation for future comparative analyses. In this article we ask: *what are the plausible contributions of SAI to Global Catastrophic Risk (GCR)*? To the best of our knowledge this is the first attempt to offer a novel, comprehensive framework for comprehending the contributions of SAI to GCR. As noted in Section 2, this is a useful and original step forward for the nascent field of studying GCRs. This is not just simply adding up SAI's potential negative impacts. It requires understanding how SAI could trigger or worsen other large-scale threats (such as nuclear warfare) or systemic risks. Understanding extreme downside risks can also help provide direction for policy and governance. The future may be hazy, yet avoiding the extreme downsides is a priority for risk management under uncertainty. To guide our investigation, we put forward a novel framework for understanding how SAI, or any other complex risk, contributes to GCR. We then use this to review and discuss the existing evidence on SAI's critical threats.

In Table 1 we provide a brief set of definitions of the key terms we use throughout this chapter.

Table 1: Definitions.

Term	**Definition**
Climate Engineering	Large-scale, deliberate interventions into the Earth system to mitigate the effects of negative impacts of climate change.[1]
Extinction Risk	A risk that could plausibly cause human extinction.
Global Catastrophic Risk (GCR)	A risk that could plausibly cause a loss in global population of 10–25%[2] and a disruption to one or more global critical systems.
Solar Radiation Management	Measures which impact the albedo of the Earth system in order to mitigate the impacts of climate change.
Stratospheric Aerosol Injection	The injection of light-reflecting chemical, such as sulphur dioxide, into the stratosphere.

Term	Definition
Systemic Risk	The ability for an individual disruption or failure to cascade into system-wide and cross-system failures[3] due to structural conditions.
Latent Risk	Risk that is dormant under one set of conditions but becomes active under another set of conditions.
Termination Shock	A large and rapid increase in warming after the cessation of SRM measures.
Buffering	A period of roughly several months following the cessation of SAI where effects of termination shock do not occur. Redeployment of SAI during this period would ensure that termination shock does not occur.[4]

Our approach makes use of a structured literature review and systems mapping exercise. We use our novel framework to structure a literature review covering studies relevant to the risks of SAI. For each area we highlight the level of evidence and uncertainty, and draw out some key implications. The nature of the risk will depend on the specifics of the geopolitical situation and the SAI intervention. We explore this through a causal-loop diagram (Figure 1) which plots out the connections between the level of risk, the amount of SAI loading, the level of international coordination and other key variables.

Note that for most of this chapter we address SAI in the abstract. The exact potential damage imposed by SAI would vary the way it is deployed. In Section 7 we discuss how the method of deployment creates different impacts. Throughout the chapter we assume a "default" deployment method of SAI to be the continuous multi-decadal global use of planes with multiple injection locations, guided by a global cooperative endeavour led by states with private sector contributions, with an overall objective to respond to global warming. Deployment "thickness" (how much warming is masked) is a particularly important variable. We flag thickness throughout our analysis. Where we discuss the risks of other potential forms of deployment we directly state so.

We proceed by outlining our framework (Section 2), before examining SAI's direct catastrophic risks (global ecological impacts; Section 3), SAI's interaction with other catastrophic hazards (Section 4), SAI's potential input to systemic risk (Section 5), and finally SAI's influence on latent risk (Section 6). We then discuss how different methods of deployment could lead to different risks and what the policy implications of our analysis are (Section 7). To avoid the critical downside risks we consider throughout the chapter, SAI governance would have to be near perfect for multiple decades.

A solution that is almost impossibly difficult to implement well, and that plausibly threatens catastrophe if implemented poorly, is not a good solution.

Whether this is preferable to climate change remains to be seen.

2. A Framework for Unravelling Global Catastrophe

There is no agreed framework for understanding the contribution of different phenomena to GCR. Most studies and reports on GCRs rely on analysing a set of large-scale "GCR-level" hazards.[5] Usual suspects include anthropogenic risks such as nuclear weapons, climate change, and more speculatively, Artificial General Intelligence,[6] biologically engineered pandemics, and natural risks such as super volcanoes and asteroids. While there have been some alternative frameworks for classifying GCRs,[7] these have yet to be widely adopted. They are also disconnected from relevant literature on systemic risk. Moreover, while they are helpful in classifying a given hazard, they do not act as aids in understanding how much a given event or system could contribute to overall levels of GCR or extinction risk.

There are several problems with the typical, hazard-centric approach. First, it is unclear how these hazards are decided on. Second, a risk is composed of hazards, vulnerabilities, exposure, and response, not just individual threats.[8] Third, the different hazards are treated as disconnected when they frequently have similar institutional drivers. Fourth, it ignores systemic risk, particularly the ability for a set of smaller, diffuse risks to scale to a global and cataclysmic level due to the fragility and interconnectedness of critical systems. Any practical framework needs to consider exposures, vulnerabilities, and drivers as well as their interlinkages.

We put forward a four-stream framework for understanding the contribution of a system or event to GCR. This rests not upon having a particular probability of occurring. Instead, we focus on what is plausible (rather than "merely possible"), consistent with our background knowledge of physical and social systems.[9] Understanding risks which are plausible, high-impact, but low- or unknown-probability is critical for robust decision-making under uncertainty.[10] For example, making decisions on the "better" worst case is central to the Maximin approach. The framework covers the hazard, vulnerability, and exposure elements of risk. Hazards are directly assessed through the first two streams of the framework, while the focus on systemic risk analyses potential vulnerabilities. Latent risk explores the often-neglected possibility of vulnerabilities that are hidden in the short term. Exposure is articulated throughout the analysis. Response (i.e. SAI governance) is discussed throughout Section 7.1.

Our four-stream model looks at direct contributions to GCR, how it could potentially trigger other high-impact risks, its contribution to systemic risk in global critical systems, and its capacity for latent risk. Across each, we also consider potential feedback loops between SAI and each stream. Our four-stream model is as follows:

1. The first stream focuses on directly catastrophic impacts. A direct contribution refers to ways in which the impacts caused by SAI could alone plausibly cause sufficient mortality and morbidity without considering wider social knock-on effects.
2. The second stream examines how SAI could interact with other high-impact hazards such as nuclear war.
3. The third investigates how SAI could contribute to and be affected by systemic risk. Systemic risk focuses on how structural conditions and multiple small stressors can lead to widespread collapse or synchronous, reinforcing failures.[11] Indeed, complex systems can undergo rapid degeneration even without large shocks. They frequently organise into critical states in which small perturbations quickly cascade into calamity.[12]
4. The final stream focuses on SAI's latent risk. Latent risk focuses on deciphering how SAI could pose threats that manifest under post-catastrophe conditions, such as in the aftermath of societal collapse.

Together, these different factors provide a comprehensive framework for comprehending how SAI could raise or lower overall levels of GCR in the world. The framework is intended to be a first step to risk comparison, in this case climate change and SAI. These streams echo the channels of risk discussed in "Climate Endgame".[13] This helps make any risk-risk comparison easier. While the framework is extensive our application is limited. Due to time and resource constraints we only explore the most well-evidenced and likely risk channels, leaving others relatively untouched, such as the impact of SAI on the likelihood of biologically engineered pandemics.

Historically, comparison between the two has been a rhetorical device to justify SAI. This is by no means a straight-forward juxtaposition since the two interact (for example, through mitigation deterrence: actors may be less open to ambitious emissions reduction if there is a "technofix" on the horizon[14]) and any analysis hinges on subjective judgements about climate sensitivity, tipping points, adaptive capacity, and the likelihood of international cooperation. There is also the issue of which precise baselines should be used for comparison:[15] what should climate change or SAI be specifically compared against? In addition, how should the two be compared? Given the high uncertainties for both climate change and SAI, is a Maximin analysis of the "better" worst case a prudent or viable approach? Given these difficulties, we do not look to provide a definitive answer or quantitative analysis. Ultimately, we are not just comparing two different sets of risks, but two separate Earth system states[16] with different winners and losers. Navigating these entangled risk analyses is an area for future analysis, but analysis that this chapter can hopefully inform.

Nonetheless, any public deliberation and democratic decisions need to rest on comparable evidence and information. Any action is bettered by risk assessment, even if it is always mired in uncertainty. This article provides an initial and incomplete basis for informing such discussions. Imperfectly mapping out risk trade-offs is preferable to sleepwalking[17] into a dangerous future.

3. Directly Catastrophic Impacts: Ecological Blowback?

Could SAI lead to directly[18] catastrophic ecological impacts? Existing studies highlight a raft of potential negative consequences. But the specific nature of these impacts, and their contributions to catastrophic outcomes, depends on the specific SAI implementation. This is an issue of high uncertainty, particularly regionally.

The projected local ecological effects of SAI are mixed and uncertain, depending on the specific analytical approach and specific SAI deployment. Monsoon areas would likely face a drop in precipitation under large scale SRM,[19] but this focuses on SRM in the abstract and may not be fully applicable to SAI. Many regions could face a seasonal under- or over-compensation in rainfall (compared to a high warming average (RCP 8.5) from 2010 to 2030, and assuming SAI is implemented to mask five degrees of warming).[20] Effects on hydrological systems would be regionally diverse and uncertain due to potential changes in nonlinear variables including surface runoff, evapotranspiration, rainfall levels, and distribution.[21] These fine-grained changes in weather could then affect vegetation. Plant communities could transform their structure, traits, and geographical range, particularly under larger swifter SAI deployments.[22] While SAI might offer salvation to climate vulnerable vegetation it will depend on deployment timing. Some communities may already be committed to at least local extinctions before SAI is deployed. SAI would likely result in ecological trade-offs with some communities benefitting and others suffering. The exact nature of these trade-offs is uncertain and needs further study.[23] The key theme here is that SAI would likely have a range of impacts on many ecological systems. But how these would play out is highly uncertain, particularly at regional scales. Impacts hinge on the inherent uncertainties within complex ecological systems, varied comparative baselines, and the specific SAI deployment.

The overall direct impacts of SAI, while uncertain, do not currently seem to constitute a catastrophic threat. Whether SAI would cause greater risks in terrestrial, freshwater, marine systems than climate change is unclear and depends on SAI's specific deployment configuration. Higher levels and swifter deployment of SAI would mean greater potential for disastrous impacts.[24] Additional considerations like seasonal[25] or hemispheric[26] deployment further affect potential impacts.

There is a paucity of research on SAI impacts,[27] particularly so for catastrophic or worst-case impacts. This has been the case for climate modelling literature in the past as well.[28] Climate modelling is often an exercise in "betting on the best case".[29] Others have noted this idealistic tendency for SAI modelling:[30] for example, limiting SAI use to only halving warming[31] or limiting SAI deployment to spring.[32] These idealised approaches in theory could reduce negative impacts associated with SAI. Yet their likelihood is questionable due to optimistic assumptions of multi-decadal international cooperation (see Section 5.2).

The possibility of dangerous ecological tail-risks depends on the level of cooling. Initial game theoretic research indicates the possibility of overcooling if SAI is pursued by uncoordinated actors.[33] Negative impacts which are projected to be relatively minor in existing studies — for example, sulphate deposition impacts on terrestrial ecosystems[34] — may become major ecological issues if SAI is deployed to far more of an extent than envisioned. Similarly, a poor choice[35] of aerosols could result in large-scale ozone depletion.[36] It is unclear whether, in these extreme cases, biophysical impacts would revert to their pre-SAI state once SAI is removed. Modelling on "worst" cases is thus critical in informing SAI's desirability. Exploring uncoordinated scenarios with the (simultaneous) use of different aerosols, different desired extents of cooling, and implementation by a small club, would all be helpful complements to existing idealised modelling scenarios.

Regardless of how developed our understanding on SAI impacts become, there will always be inherent uncertainty. When dealing with a complex system like the climate there is always the chance that a black swan is lurking in the dark.

Some commentators have downplayed the potential of unknown impacts due to the availability of historical analogues, namely historically severe volcanic eruptions.[37] Improvements in modelling, a gradual implementation, and a cessation if unacceptable negative impacts are found could also lessen the likelihood of an unforeseen catastrophic tipping point.

None of these reasons are causes for comfort. Modelling, regardless of improvements, may simply be incapable of capturing rare tipping-points and is not intended to accurately predict or foresee non-rational political dynamics.[38] In addition, a gradual rational phase-in and phase-out relies on optimal governance conditions. Overly rapid deployment due to "free-driving"[39,40] or overly slow phase-out due to technological or infrastructural lock-in[41] are entirely plausible. Moreover, SAI impacts may also not follow the pathway of historical analogues. The core rationale of SAI is to manufacture the cooling effect of a volcanic eruption in a "safe" manner, not replicate volcanic processes. Deviance from historical analogues is especially a possibility if the choice or mix of aerosol is radically different. This is particularly the case since climate change and human-pressures are already pushing ecological systems into novel states.[42] SAI would push systems into further novel states that make unseen ecological responses likely.[43]

Our understanding of both Earth systems and the likely contours of deployment are too weak for us to rule out a potentially catastrophic form of ecological blow-back. For now, the literature points to SAI having numerous impacts. But none seem remotely capable of being a GCR, particularly if SAI deployment were limited. Nonetheless, the spectre of an unforeseen tipping point in the Earth's climatic system remains.

4. Interactions With Other Global Catastrophic Hazards (GCHs)

The impacts of SAI, or any other catastrophic risk, should not be assessed in isolation.[44] Different catastrophic hazards[45] have interactions. One could potentially trigger another and/or worsen its effects. Climate hazards, for example, have been shown to compromise governments' ability to provide effective responses to COVID-19.[46] The potential for one global shock to ignite and amplify another has previously been dubbed "double-catastrophes".[47] Baum, Maher Jr. and Haqq-Misra (2013) suggest that this could be the case if nuclear war or a pandemic were to disrupt an SAI system, leading to abrupt termination shock. GCHs which are simply a matter of probability, like extreme space weather or a volcanic eruption, may also coincide through pure bad luck.

In this section we consider both a broader array of hazards and how SAI could trigger and interact with them. This will not be an exhaustive comparative analysis of all possible GCHs. Instead, we focus on hazards that have clearly established causal relationships, relatively well-developed literatures, and some empirical track record of their impacts. Our analysis suggests that the possibility of SAI sparking other GCHs are tenuous. SAI could only plausibly contribute to large-scale conflict and potentially nuclear war. The possibilities of SAI exacerbating other GCHs are more concerning. SAI has the worrying ability to significantly heighten the impacts and mortality of any global catastrophe due to termination shock.

4.1 Volcanic eruption

A large volcanic eruption would demand rapid SAI adjustments. While severe overcooling seems unlikely (the cooling of SAI and volcanic winter are not additive),[48] SAI should be rapidly scaled down in a matter of weeks.[49] Laakso et al. (2016) assume a relatively thick SAI injection (offsetting roughly a doubling of carbon dioxide from preindustrial levels). The prudent course of action for thinner SAI is unclear. However, the SAI adjustment in a volcanic future is not simply one of scale down. SAI injection may need to increase in the opposite hemisphere to the volcanic eruption to ensure a more uniform global temperature[50] (a high temperature variance across hemispheres can have severe adverse impacts on precipitation and drought dynamics).

Adjusting the SAI level may seem straight-forward but depends on an informed, rapid political response. There are reasons to doubt this would be forthcoming. First, the technical demands may prove too much for cumbersome domestic and multilateral politics. These include potentially politically vexing dilemmas over the balance between scaling SAI up and down on different hemispheres, whether to inject SAI at new locations or "thicken" existing deployments,[51] or whether SAI should be scaled down at all. A second and novel addition is that a volcanic eruption would not solely affect temperature. Many pinch points of global supply systems are near active volcanic areas. Even modest volcanic eruptions could lead to disruption and catastrophic economic system collapse.[52] The difficulty of coordinating regional SAI adjustments would be compounded by sub-optimally functioning supply systems and general economic and political chaos.

While the interactions between a volcanic eruption and SAI currently seem to have only modest direct contributions to catastrophic risk, the highly political decisions of a volcanic-SAI world may lead to political ruptures and ineffective SAI governance.

4.2 Space weather

Solar flares, coronal mass ejections, and associated solar radiation and geomagnetic storms, can lead to widespread damage to terrestrial, avionic, and space infrastructure. The fear for SAI is that a "black sky" event could disrupt and knock out critical SAI infrastructure. Yet there have been no attempts thus far to investigate SAI-space weather interactions. We examine SAI interactions with an Earth-bound space weather event roughly on par or worse than the 1859 Carrington Event — the benchmark for extreme space weather events.[53] A current-day Carrington Event would likely lead to widespread electrical failure and disruption for multiple months at minimum, potentially years.[54]

Extreme solar events are difficult to accurately and timely forecast. They are essentially random events[55] which provide little forewarning. Solar radiation can travel at such high speeds that an extreme coronal mass ejection would likely reach Earth in less than a day. Other radiation and energised particles travel at or close to lightspeed — eight minutes to reach Earth. Even with the earliest detection possible there would be little response time.[56] It would be a late flinch to an oncoming blow.

The impacts of extreme space weather events are vast. Aviation, satellite, and general electronic infrastructure are especially vulnerable. Energised particles can affect memory cells — for example, changing a bit from a 1 to 0 and vice versa — that lead to erroneous commands or overall hardware failure.[57] Global navigation and communication systems would experience disruption and downtime that could last several months (alternative navigation systems, like the US Alternate Position Navigation and Time programme, may still be affected by electrical damage).[58] Aircraft crew would have greatly limited airtime due to limits of safe radiation exposure.[59] Flights at higher altitudes and closer proximity to the Earth's poles would be unlikely to continue.[60] The use of automated aircraft would be compromised by widespread electrical and avionic damage. Especially alarming is that SAI would likely depend on vulnerable aviation, satellite, and general electronic infrastructure for deployment, monitoring, impact attribution determination, calibration, and modulation.

Impacts of space weather events are not limited to human infrastructure. Substantially increased UV output can influence the Northern Hemisphere jet stream, ozone production (and ozone UV absorption and warming), and precipitation patterns.[61] These systems, particularly precipitation, are the same systems that SAI is likely to greatly affect. Interaction between these impacts is currently unclear.

These disruptions appear enough to halt even a robust SAI system. Even with high uncertainties of potential infrastructural impacts[62] and the nature of the event itself,[63] the limited evidence so far indicates that SAI infrastructure would be vulnerable and exposed to damage, thus leading to termination shock if SAI was sufficiently thick (see Section 6). In the aftermath of an extreme space weather event, continued implementation or preservation of SAI infrastructure would have to compete for limited government attention. Damage would be widespread and international — ranging from railway failure[64] to power failure[65] to failure of satellite infrastructure.[66] Governments and resources would be stretched thin and SAI reimplementation may be neglected. An extreme space weather event could lead to severe economic and infrastructural shocks[67] that make continued SAI deployment infeasible. At worst, widespread power failures could lead to ripple effects across food, health, and transport systems that extend recovery time potentially into decades, driving modern societies back to a more fractured pre-electronic state.[68] It is unclear how SAI, with its high technical and information demands,[69] could continue under these conditions. Troublingly, mitigation options are currently limited and highly depend on future (but relatively well-known) scientific and engineering solutions.[70] Considering the speed of space weather events, SAI infrastructure would have to be built to be resilient (with technology which does not currently exist) from the offset.

SAI is ultimately highly vulnerable to extreme space weather events. Widespread electrical damage would compromise SAI redeployment, making a termination shock highly likely and worsening the already catastrophic impacts of an extreme space weather event.

4.3 Nuclear weaponry

SAI would likely worsen any nuclear winter and our recovery from it. A nuclear war could occur due to either an accidental strike leading to escalation, or a full-blown exchange. Even a relatively smaller conflict between Pakistan and India would have global ramifications. The background risk of incidental or inadvertent nuclear deployment is present unless there is total nuclear disarmament.[71] In addition to nuclear winter, the physical blast, ionising radiation, and electromagnetic pulse (EMP) would all contribute to widespread and severe damage of electronic infrastructure,[72] including SAI infrastructure. Indeed, EMPs are similar in effect to the "black-sky" events discussed in Section 4.2. This leads to two key concerns. The first is the combination of SAI cooling with nuclear winter conditions, the second is the grim mixture of nuclear cooling combined with termination shock.

The combination of SAI's existing cooling and additional nuclear winter would likely lead to short term overcooling, followed by medium- or long-term overheating due to termination shock.[73] It could be global frost followed by global furnace. Alternatively, there may be the potential for SAI and nuclear winter layering to spark non-linear or unexpected cooling effects. This is an area that justifies further study. There is modelling on the impacts of a nuclear detonation, comparison of nuclear and climate threats via the "climate-nuclear nexus" (Scheffran et al., 2016), and modelling on the impacts of SAI deployment and termination shock. Yet so far nothing integrates these two separate bodies of knowledge. The oversight is interesting given the entangled histories of climate science and nuclear weapons research.[74] For now, the interactions between nuclear winter and SAI remain neglected and our analysis here is hence provisional. In any case, such rapid swings in global temperature would be unprecedented for the Earth system and humanity.

A key question is whether a disrupted SAI system could be revived during nuclear winter to prevent a termination shock summer, and whether SAI was masking sufficient warming for termination shock to occur (see Section 6). But there also are reasons to believe that the re-establishment of an SAI system would not be able to occur during the buffer period in the wake of a nuclear cataclysm. First, technological damage may be so severe that timely deployment is impossible. Backup infrastructure like aircraft (and associated supporting infrastructure such as air traffic control) may be damaged beyond repair or be grounded for security purposes. Second, political and policy attention would likely be focused on other post nuclear issues, such as disaster recovery and the creation of alternative food systems. As with other disasters, governments would be stretched thin and may prioritise these more short-term issues. Lastly, a post-nuclear world would likely exhibit a lack of international cohesion that is seen as an enabling condition for effective SAI.[75] Discussions over SAI have already been deadlocked.[76] It seems unlikely that a world of post-conflict lessened trust would be more conducive to speedy decision-making. Different countries may drop out of implementation, further complicating SAI deployment configurations, possible regional impacts, and concordant policy responses. Disagreement over resource allocation is likely to arise, as is the case for many disaster recoveries.[77]

The presence of thick SAI greatly increases the potential consequences of nuclear warfare, and vice versa. The rapid temperature swings involved with a nuclear winter and termination shock summer would likely lead to ecological disaster, and a chaotic post-nuclear world would not likely reimplement SAI in a timely sensible manner.

4.4 Pandemics

A pandemic that reaches the level of a GCR could be enough of an economic or population shock to sever an SAI system.[78] Whether the system could be reactivated during the buffer period would depend on both the severity as well as the length of the pandemic. COVID-19 provides a chilling reminder that states are not rational nor necessarily cooperative during a disease disaster. COVID-19, a far cry from being a GCR, has spawned fragmented responses and cases of both vaccine nationalism and vaccine diplomacy. Such multilateral behaviour does not engender confidence that a pandemic with a significantly higher mortality rate would lead to survivors coolly and collectively reactivating an SAI system whilst dealing with the outbreak. Other issues, like keeping healthcare systems afloat, would likely be an overwhelming priority. With resources and capacity stretched thin, SAI may be neglected. A pandemic would be a severe shock to political and economic systems that may preclude continued SAI use, not least rational, well-governed, well-resourced SAI use. Whether this risks termination shock depends on the amount of warming masked.

There are also reasons (albeit speculative) to believe that SAI could contribute to a pandemic. SAI induced temperature changes and uncertain regional climatic effects can alter disease transmissions.[79] This could in turn affect pandemic dynamics. As with general ecosystem impacts (Section 3), a larger and quicker SAI deployment can be expected to have more severe impacts. Critical nodes in urban and health systems may become exposed to diseases that are beyond typical immunity or resistance (see Section 5.3 more on SAI-health interactions). This could be the spark for a pandemic spread, particularly if decision-makers are unprepared to make early and rapid response measures. However, the most worrying (but thus far neglected) concern would be effects on animal populations. Similar concerns of low or lapsed immunity or resistance would apply to animal populations and new disease vectors. But animal populations would lack similar healthcare systems to keep disease spread at bay. Many contemporary pandemics have resulted from cross-species spillover,[80] including the 2009 Swine Flu Pandemic from pigs and birds, and the 2013–2016 Ebola Epidemic and COVID-19 Pandemic from bats. Altered animal disease dynamics, particularly those stemming from unpredictable regional SAI impacts, may increase the frequency and severity of future pandemics.

5. The Systemic Risks of Climate Engineering

Both previous societal collapses and disasters in the modern world are marked more by the accumulation of many stresses leading to failure, rather than single abrupt shocks destroying systems.[81] Seemingly modest stressors can cascade to catastrophe. This section analyses the potential of SAI to create and be impacted by biophysical and political stresses which contribute to global systemic risk.

The world currently exists in a deeply interconnected, and increasingly homogenous state which is prone to systemic risk.[82] One ship blocking the Suez Canal in March 2021 led to losses of roughly $6–10 billion.[83] More serious stressors could lead to far more severe consequences. The economic and political state of the world would be central in determining whether risk cascades. It is unclear how SAI could or would adjust the structure of the globalised economy. Hence, instead we focus on a few critical systems that SAI might be expected to impact and where there have been initial attempts to gather evidence: agriculture, health, and international politics.[84] SAI would likely not alter any of these system structures, but would rather aggravate existing systemic vulnerabilities.

5.1 Agriculture

SAI's effects on temperature and precipitation distributions would likely affect agricultural systems. The precise nature of these impacts are unclear.[85] For example, some studies have shown that the low temperature, high carbon dioxide environment of a SRM deployment might increase yields: maize yields may increase in China,[86] as could overall global yields of maize, wheat, and rice.[87] On the other hand, solar dimming might reduce yields of groundnut in India[88] or offset benefits of reduced temperature.[89] These effects would all further differ across crop and area. The differing approaches to analysis (Xia et al. (2014) focus on SRM to offset a 1% increase in carbon dioxide from preindustrial levels for 50 years, whereas Pongratz et al. (2012) focus on SAI masking carbon dioxide concentrations of 800 ppm) as well as use of outdated equatorial injection in these studies (see Section 7.1) make clear conclusions difficult to discern. The main point is that SAI would affect agriculture, but the precise impacts are unknown.

Regardless, the sensitivity of these key staple crops alone is a cause for concern. Small variations in yields of staple crops could induce disproportionate price fluctuations and cascades into socio-political violence, particularly in areas with political instability and weaker governance.[90] Additional uncertainties with attribution between SAI and agricultural yields could compound potential political difficulties.

Even in the case that SAI provides agricultural benefits, these are likely to be marginal if other issues affecting agricultural productivity, such as habitat loss and soil degradation, continue unabated.[91] An SAI high carbon dioxide, low sunlight world would also require additional adaptation on the part of agricultural actors. This does not look likely given agricultural adaptation to climate change has so far only been modest.[92] Large-scale changes in yield and precipitation are likely to create at least short-term food insecurity. There is evidence that existing population density and economic growth are closely tied to the existing climate niche. The narrow climatic envelope of ~13 °C has provided beneficial environmental conditions within which most humans and societies have tended to historically cluster.[93] Our agricultural systems almost certainly are similarly tied to this niche, and any sudden change at a global level is likely to affect short-term yields and prices.

5.2 Politics

SAI could feasibly spark conflict and instability. There are already some emerging empirical links between food price shocks and socio-political violence. Moreover, the very act of undertaking SAI could be grounds for dispute. States may look to develop their own SAI capabilities before others do, creating more extensive backup infrastructure to avoid dependencies on others, or even construct counter-SAI capabilities.[94] Existing political order may become undone by SAI.[95] A novel and interesting example could be high historical emitters like the US using the Common but Differentiated Responsibilities and Respective Capabilities principle as an instrument to assert SAI control or leadership ("we are mostly responsible for climate change, therefore it is 'just' that we lead the response"). Manipulation of the climate could become a new frontier for political conflict or even warfare. Different cross-boundary impacts on different regions would create large sets of winners and losers, alongside questions of attribution[96] and compensation. Whether such disputes could snowball into conflict is beyond prediction. Nonetheless, it is reasonable to say that unless enacted as altruistic, cooperative endeavour over multiple decades, the project of SAI would load further pressure onto existing international tensions. But even in the most altruistic cooperative scenarios, there may still be sub-national tensions in and/or between "donor" and "recipient" populations.

There is also the possibility that politics would worsen SAI. SAI and politics is a two-way street. Political conflict can cascade to affect SAI deployment and its impacts. Previous studies have made a compelling case that the direct weaponisation of SAI is unlikely.[97] High-impact uncertainties, management difficulty, low precision, and preferable alternative weaponry make SAI an unappealing instrument in state arsenals. However, this does not mean that SAI has limited military use. SAI may not have usefulness as a direct weapon, but can function as a support system or a threat. Indeed, early attempts at cloud seeding were used by the US military during the Vietnam War as a tactical weapon to extend the Monsoon Season and disrupt North Vietnamese supply lines (Operation Popeye).

Another avenue for political dynamics to worsen a SAI deployment, and that has received relatively little attention, is via cyberwarfare. In May 2021, a ransomware cyber-attack forced a US fuel pipeline out of service. A $5 million ransom was paid to restore service.[98] As a globally critical (and potentially highly politicised) piece of infrastructure, SAI would likely be a target for private or state actors. SAI deployment dependent on any software or advanced algorithmic system,[99]which is likely given the high technological and informational demands of deployment,[100] would be vulnerable to cyberattack.

Cyberattacks do not need to come from external forces. For instance, the notorious 2000 Maroochy Cyberattack was from a disgruntled ex-employee.[101] SAI would likely depend on a large workforce and have numerous reasons for controversy.

These political dynamics would have decades to play out. A cooperative and benevolent deployment of SAI could crumble into chaos with a change in actor preferences (or vice versa). Politics and its broader conditions are likely to change substantially over coming decades. Interactions between future geopolitics, warming and emissions, and technology are all nigh impossible to predict or even foresee,[102] but would be of critical importance to SAI and its governance. Relying on one set of optimal political assumptions would be greatly unwise.

5.3 Health

SAI could negatively impact human health by both changing disease vectors and range (and therefore pandemics, see Section 4.4), and by undermining existing health system infrastructure. The regional variations of SAI's impacts on temperature and other ecological factors would likely affect disease transmissions. SAI-induced reductions in monsoon rainfall may increase cholera risk,[103] and temperature changes can affect transmission of vector borne diseases like malaria.[104] Yet such health impacts are chronically understudied: currently only four papers focus on the health impacts of SAI.[105] The lack of coverage is significant since these studies have critical limitations, namely an assumption of equatorial injection (see Section 7.1). The impacts of other forms of deployment are largely unknown. Similarly, there is little research on the health impacts of exposure to SAI aerosols,[106] and the few quantitative assessments of mortality related to air quality and changes in UV exposure carry significant uncertainty.[107]

Despite these limitations, the research to date does point towards potential dangers. Alterations of disease transmission are especially important because diseases may reach populations which have lapsed or little immunity or resistance,[108] or may have relatively weak or vulnerable public health systems. These critical nodes in health and urban systems, which otherwise would be less exposed, may amplify health risks and impacts: an epidemic may be amplified to become a pandemic (Section 4.4). The uncertainty of SAI's potential deployment configurations, associated impacts, and state of existing health systems means that early identification of different critical nodes would likely be difficult and insufficient. Overall, systemic effects between health and SAI currently seem modest and carry high uncertainty. However, they are not negligible.

6. Latent Risk and SAI

Latent risk refers to risks that are dormant, but could become manifest during times of heightened societal vulnerability. The most obvious example would be the additional risks that arise in the aftermath of a collapse (widespread, significant, and enduring loss of life, political organisation and economic capital) or another global catastrophe, for example violent conflict over food and water. Latent risks are particularly important as they can provide one tangible way in which recovery from global shocks could be undermined and spiral towards extinction risk.[109] We have already dealt with these partly in Sections 3–5. In short, latent risk is perhaps the largest risk factor for SAI. SAI changes the nature of climate risk by making the "likely" outcomes less severe, but making "less likely" (or "fat-tail") outcomes substantially more severe. The risk of termination shock thickens the tail. Large amounts of SAI loading could create a precarious condition in which any sufficiently large global shock is likely to be compounded by a tumultuous termination shock. It is in these worst cases where SAI becomes clearly worse than worst case climate change.

While there is subjectivity as to what a "threshold" for termination shock would be, Parker and Irvine (2018) suggest a SAI cooling threshold of around 0.3 degrees, implying a termination of at least 0.15 degrees warming per decade. Kosugi (2013)[110] puts the termination threshold at 0.2 degrees, implying an SAI cooling threshold of around 0.4 degrees.

The speed of termination shock depends on the form of SRM. SAI has a half-life of approximately eight months (approximately half the levels of coolants would still be present after eight months) and warming would still take several years to reach its unmitigated levels.[111] Depending on the amount of warming masked, SAI has a distinctly high latent risk due to termination shock. A temperature rise of six degrees in the space of centuries would be an order of magnitude faster than the warming experienced during the Great Permian Dying.[112] If experienced in a period of decades, it would be an order of magnitude faster still. Current warming rates are geologically unprecedented; this speed would be chillingly rapid.

Critics have framed termination shock as an overblown problem for numerous reasons. These include that countries are unlikely to willingly reverse SAI, that there would be a sufficient buffer period to resume SAI, and it is unlikely to be hiding a large amount of warming.[113] These all seem to align with the inclination for both modelling and analysis of geoengineering to focus on the "best case": that there would be sufficient cooperative governance and deployment of SAI, that there would be rational responses to any system lapse or shock, and that SAI would be used to only shave-off a small amount of warming.[114] Yet, SAI is widely portrayed as an emergency response: it is most likely to be used in a worst-case high warming scenario, not a best-case limited warming one. Moreover, the likelihood of high-end warming, governance fragmentation, or another GCR occurring are all disarmingly large.

The likelihood of a catastrophe curtailing SAI efforts and causing termination shock is usually dismissed as very low. This is likely mistaken. We have covered some of these catastrophes in Section 4. While there is considerable uncertainty, the likelihood of a GCR in the coming centuries does not appear to be vanishing. Estimates for a large-scale space weather event over the next decade or so range from 0.46%[115] to 20.3%.[116] Estimates of the probability of nuclear war are few and vary, but one model of inadvertent conflict between the US and Russia using historical data put it at 0.9% per year.[117] SAI could also be slowly scaled back as mitigation and CDR efforts increase.[118] But this would likely require multiple additional decades which would (assuming no mitigation of other global threats) incur a higher likelihood of another catastrophe striking.

A more compelling retort is that SAI could be reintroduced within years at a reasonable cost. Some have suggested that given that SAI could be run at >1% of the GDP of the G20 and hence even losses of 75% of GDP (an unprecedented economic disaster) would be insufficient to keep an SAI system deactivated.[119] Such analysis overstates the coherence and rationality of states responding to crisis. The value of an extra three billion doses of COVID-19 vaccines would provide benefits of $17.4 trillion, at a cost of around $18–120 billion.[120] Yet vaccine production remains chronically low. Even in far less dire circumstances we can clearly not trust decision-makers to take the optimal course of action.

SAI can be seen as one vast project to make the climate system more tightly coupled and synchronised with the global economic system. From a resilience perspective, such efforts are a liability. It makes it far more likely that the failure of one system will spill over into another, sparking non-linear feedback loops that result in "synchronous failures".[121] There are of course ways to make such complex engineering systems more resilient and robust, namely via backups and redundancies. However, current economic incentives for efficiency (particularly via cost reduction), mean that strong redundancies are rarely in place. SAI redundancies specifically are likely to be expensive and thus inconsistently implemented.[122] In any case, it is unclear what redundancies would be effective at making an SAI system catastrophe-proof. Making SAI resilient to natural disasters or terrorist attacks seem relatively straight-forward,[123] but the same cannot be said of a true global catastrophe.

The inherent unknowns of highly complex technological systems also contribute to the possibility of termination shock. Highly complex systems, like SAI would be, are prone to "Normal Accidents".[124] Large-scale accidents and disruptions are to be expected in sufficiently complex and tightly coupled systems. Unforeseen technological failures are simply a fact of life.

While latent risk is a genuine concern, it is a danger for only the greatest threats on the horizon and the "thickest" SAI deployments. A true, dramatic, global calamity would be needed to both disable an SAI system masking a large amount of warming and keep countries either preoccupied or incapable of reinstating it for several years. In this context, the risk comparison between SAI and climate change becomes clearer. SAI's worst case outcomes through severe termination shock are worse than the worst cases of climate change.

In Table 2 we summarise our analysis of direct impacts, GCH interactions, systemic risks, and latent risk.

Table 2: Summary of SAI's direct impacts, GCH interactions, systemic risks, and latent risk.

Contribution to catastrophic risk	**Type of contribution**	**Nature of evidence base and uncertainty**	**Dependency on mode of deployment**
Destabilising ecological systems	Ecological Blowback	Limited evidence base and high uncertainty. Lack of study on worst-case ecological impacts. High regional variations and uncertainty.	High dependency. Direct SAI impacts vary with thickness and other injection variations.
Volcanic eruption leading to political SAI difficulties	GCH Interaction	Limited evidence base. Study of SAI interactions with a volcanic eruption is limited.	Medium dependency. Dependent on potential SAI supply routes, but also external political dynamics.
Extreme space weather event damaging SAI and global electronic and power infrastructure.	GCH Interaction	Limited evidence base. No specific study of the impact of an extreme space weather event on SAI. Varying estimations of space weather probability.	Medium dependency. External SAI support systems are vulnerable. Thick SAI leads to more severe termination shock.
SAI-nuclear winter overcooling or nuclear frost-termination furnace.	GCH Interaction	Limited evidence base. Existing study on nuclear winter effects and SAI effects, but nothing that studies interactions between both.	Medium dependency. Dependent on external political dynamics. Thick SAI leads to more severe termination shock.
Political instability of a post-nuclear world on SAI redeployment	GCH Interaction	Limited evidence base. Existing study on post-nuclear politics and SAI politics, but nothing that studies interactions between both.	Low dependency. Dependent on external political dynamics.

Contribution to catastrophic risk	**Type of contribution**	**Nature of evidence base and uncertainty**	**Dependency on mode of deployment**
Pandemic leading to population or economic losses that make continued SAI infeasible.	GCH Interaction	Limited evidence base. Limited study of SAI-health intersections and no study of SAI-pandemic interactions.	Medium dependency. External pandemic, economic, and political factors are critical drivers. But thick SAI leads to more severe termination shock.
SAI weakening agricultural systems.	Systemic Risk	Limited evidence base and high uncertainty. Little study of SAI's agricultural impacts and high regional variance is likely.	High dependency. SAI's agricultural impacts are highly dependent on deployment configuration.
SAI sparking political conflict	Systemic Risk	Initial study of the political dimensions of SAI, but high uncertainty of how these political effects would play out.	Low dependency. Dependent on external political dynamics.
Political dynamics that compromise SAI safety	Systemic Risk	Low uncertainty that international geopolitics over a multi-decadal timescale is not ideal for optimum SAI governance. High uncertainty and limited evidence base as to how specifically this would play out.	Medium dependency. Political instability and conflict is core to the multilateral system, but nature of uneven SAI impacts and differing objectives contribute to political instability.

Contribution to catastrophic risk	Type of contribution	Nature of evidence base and uncertainty	Dependency on mode of deployment
SAI affecting disease transmissions	Systemic Risk	Limited evidence base and high uncertainty. Limited study of SAI-health intersections and highly dependent on external urban and health policy dynamics.	Medium dependency. Thicker SAI more likely to affect disease dynamics. But external pandemic, economic, and political factors are primary drivers.
Termination shock	Latent Risk	Limited evidence base and high regional uncertainty of precise termination shock impacts. But low uncertainty that termination shock would be catastrophic.	High dependency. Thick SAI leads to more severe termination shock.

7. Discussion: Building the Policy Boundaries for Climate Engineering

7.1 The means of deployment

Our analysis thus far has assumed a "default" deployment of optimal conditions of a global material approach to mitigate climate change. This is not necessarily the most likely scenario and the means of deployment and context will dramatically impact SAI's catastrophic risk profile. One of the critical variables to consider is the overall objective of a SAI deployment.

There are multiple potential objectives of SAI deployment, ranging from temperature reduction (of different extents), precipitation impact management, to biodiversity conservation.[125] These objectives will also depend on existing emissions reduction policies. There are also multiple potential "design" options for deployment configuration,[126] ranging from deployment timing, extent, placement, to aerosol selection.[127] The extent of cooling for example not only depends on how much aerosol is released, but the height of injection in atmosphere (lower stratosphere injection produces more cooling).[128] Much of the existing study on SAI assumes injection along the equator.[129] Equatorial injection is the most efficient if the only deployment objective is to maximise Earth's *overall* cooling. However, this would lead to high variance in temperature distributions, namely overcooling of the tropics and undercooling of poles.[130] Impacts discussed in Section 3 also can change with a non-equatorial injection — Arctic SAI, for instance, would have less of an effect on Monsoon precipitation.[131] Across all these there are key caveats. Neatly framed and optimised objectives found in modelling will not necessarily be reflected in messy and contested real life preferences, nor will SAI necessarily perfectly result in desired "design" outcomes.[132]

It is also important to consider that SAI may not be used solely to respond to climate change. The multiplicity of potential SAI goals opens the door to hidden agendas,[133] self-interest, and misuse. In addition, even if SAI is deployed under an idealistic scenario of climate altruism, there is no guarantee that this would persist. Considering that political preferences are unlikely to remain static over decadal timescales (see Section 5), SAI functions may slowly "creep"[134] into currently unknown possibilities of misuse. Such "Function Creep" and potential misuse are highly understudied in current SAI literature.[135]

Predicting or even foreseeing potential future SAI functions will forever be mired in uncertainty. This is part of why Function Creep is such a difficult policy problem. Initial study in this area for SAI highlights the potential to use SAI to "optimise" or create "designer climates".[136] Actors may, for example, advocate for deployment configurations that lead to more favourable conditions for critical staple crops, especially in response to warming impacts. These decisions may be the product of misjudgement or misinformation on SAI's causal nature. SAI may also be used to justify the continued existence of fossil fuel industries. This is a potential adverse incentive core to the "moral hazard" problem.[137] This could even create a new atmospheric political economy. The fossil fuel industry and other vested interests benefiting from the SAI system would have incentives to both use it as a way to slow decarbonisation and perhaps even thicken SAI deployment over time. This would heighten latent risk. Assuming SAI as a benign climate change response is unwisely narrow. Other more sinister SAI uses, whether purposeful or inadvertent, are critical determinants of SAI's desirability.

There are many more SAI deployment options that are not currently well captured in extant governance literature. SAI risks for example take a drastically different form if Artificial Intelligence (AI) is one of the central aspects of deployment design. With the vast amounts of information feedback and constant operational adjustments required,[138] an advanced deep reinforcement learning system may be used to manage SAI deployment.[139] This would introduce a raft of new issues: for instance "black box" opacity of decision processes[140] or inappropriate generalisations of incomplete data.[141]

Given the high variance of potential SAI objectives and potential deployment configurations, a highly political, uncoordinated, and decentralised[142] "Wild West" deployment scenario, with unclear direct impacts, is possible. States and private sector actors are not likely to find agreement on a single defined "set" of objectives, how they should be prioritised, and how these objectives should manifest in deployment configuration. These intensely political and self-interest driven considerations are likely key determinants of SAI deployment impacts, and should be priority areas for future governance research.

The means of deployment for other GCHs also affect SAI risks. An intentional weaponised pandemic may *intentionally* leverage SAI dynamics, like changes in disease transmission via changes in temperature distribution, to target critical nodes in health and urban systems. Such potential for now is speculative, but ultimately plausible.

7.2 Interconnections

Our analysis has focused on individual pathways for SAI to contribute to GCR. However, none of these are mutually exclusive. Each of the four steams overlap and feed into the same waterway. For instance, uni- or mini-lateral deployment of SRM systems could be driven by geopolitical distrust and conflict. This would likely be a world in which other GCHs are more likely, SAI deployment is less coordinated and damaging, critical systems are less resilient, and the world is less likely to quickly and effectively deal with a termination shock.

There is also an important intersection between systemic and latent risk. Most mechanisms that increase systemic risk will tend to raise latent risk as well. For instance, just in time delivery systems and tightly coupled systems with few back-ups, while efficient, are both more susceptible to shocks and can impede recovery. In Figure 1 we provide one brief attempt to map some of the linkages between different risks and factors in SAI deployment. More interconnected systems mean a higher chance of synchronous failures, and SAI is likely to be a highly interconnected system.

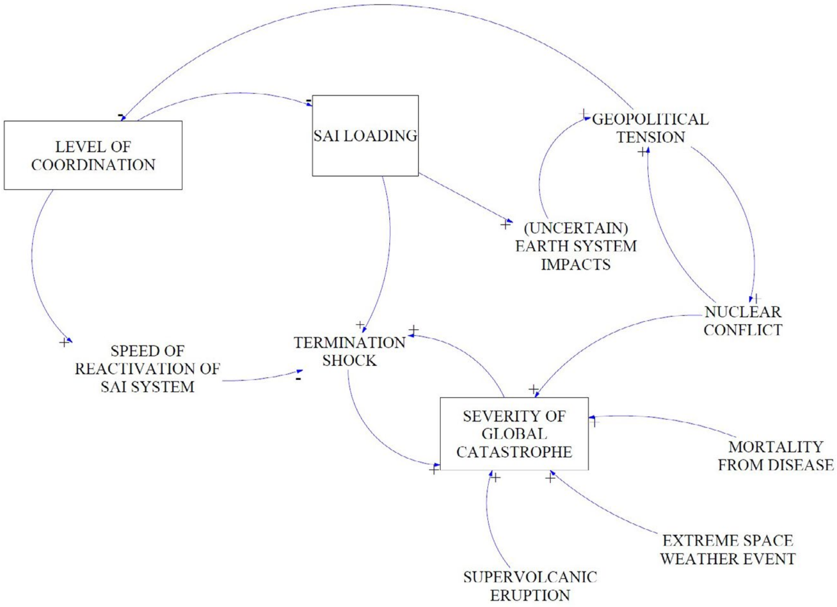


Fig. 1: The SAI-GCR System.[143]

7.3 Building the Policy Boundaries

Analysis of catastrophic downside risks can help illuminate the contours of what "effective" SAI governance would do. This is a useful complement to the policy literature that has focused mostly on structure and architecture.[144] We add to the knowledge on policy *instruments* by providing further detail on policy *approach*.

To effectively mitigate against the (limited) number of threats and systemic risks outlined in this chapter, SAI governance would have to be wide ranging, robust, and persist over decades. SAI and its backup infrastructure would need to be built to be resilient to extreme space weather or nuclear EMP events. Effective SAI governance is also not limited to SAI itself, but encompasses other policy areas like health, agriculture, AI, and energy. Ensuring ambitious emissions reductions and greenhouse gas removals would be needed to ensure SAI did not continue indefinitely. Effective SAI governance would also prevent future misuse and balance shifting preferences and multiple deployment goals in a just and inclusive manner. Governance arrangements to ensure SAI deployment or reimplementation in the wake of a major shock like a recession, pandemic, or nuclear attack would also be necessary. These would all be in addition to the herculean technical informational demands necessary for SAI deployment, which alone may be a larger undertaking than an IPCC report.[145] This optimistically assumes that SAI's climatic outputs can be clearly and cleanly measured, that there would be widespread international capacity for effective monitoring,[146] and that this monitoring would also be resilient to critical shocks. Substantial advancements in climate science and observation, as well as additional international capacity building for monitoring and transparency, would be needed. All of these would then have to be maintained over the course of decades.

These altogether represent an incredibly challenging governance task. The lack of success of the climate regime, with its similarly intense political and wide-ranging nature, does not inspire confidence in the feasibility of wide-ranging and long-term governance for an issue as political as SAI. Basic discussion over climate engineering as a whole has been stymied under the UN Environment Assembly.[147] Future and more consequential SAI debate will be subject to more severe political hurdles. Even less complex and smaller scale governance arrangements, like COVID-19 mask mandates, have been mired in politicisation and competitive dynamics. A "mask" over the Earth, and its associated governance considerations, would face even tougher political challenges to effective implementation.

What would happen if a SAI deployment went ahead without these governance safeguards? It could very well be the case that agricultural or health impacts of SAI are limited or even positive. A disaster like an extreme space weather event may not happen in the decades where SAI is implemented. The stars of international politics could align and allow for a smooth SAI implementation and cessation.

There is indeed no guarantee that the catastrophic pathways outlined in this chapter will materialise. But *if* they do, they would likely result in severe and cascading consequences. SAI has many extreme downside risks. “Imperfect” SAI governance can be compared to living without health insurance. The extra safeguards and protection aren’t strictly necessary…until something goes wrong. Given what we know about the instability of international geopolitics, SAI with imperfect SAI governance puts the world in a precarious position and introduces a climatic Sword of Damocles. The ultimate question becomes: are we willing to bet *the climate* that no catastrophe or systemic cascade will trigger SAI’s downside potential over the coming decades?

In a world of imperfect safeguards, two interconnected options are available to alleviate catastrophic risks. The first option is thinner SAI deployment. Thin SAI has a lower risk of catastrophic termination shock, thus posing less of a threat even if triggered by another calamity or systemic cascade. The second option is to ensure diversity in the overall climate engineering portfolio. Reducing reliance on SAI would better allow for a thinner SAI deployment. Other climate engineering approaches, particularly those which are less technology based, would also not necessarily share the same vulnerabilities as SAI. Trees for instance are not vulnerable to extreme space weather. These would reduce the potential of an SAI termination, but ultimately would not completely remedy the political complications SAI would create.

There seem to be three major[148] pathways moving forward. The first is living in a highly vulnerable scenario of imperfect SAI governance — the “Damocles Pathway”. This is clearly undesirable. The second is living with well-governed SAI that will not exceed policy boundaries of catastrophe — the “Miracle Pathway”. This seems infeasible. The final middle ground is to accept the inevitably imperfect contours of SAI governance, but greatly limit the extent of SAI deployment — the “Limited Pathway”. But this again would rely on robust and resilient governance and is still vulnerable to geopolitical shocks. SAI may by thickened or thinned along changing political tides.

A core conclusion here is that there is little use in asking whether SAI is a GCR or not. It depends on the level of loading and wider geopolitical landscape. All risks, especially latent risk, will increase with greater loadings and political conflict. This is a critical insight for the wider study of GCRs. A risk cannot be judged in a vacuum. Its severity will inevitably be determined by the scenario and system in which it unfolds.

8. Conclusion: The Frying Pan and the Flame

We map the different contributions of SAI to Global Catastrophic Risk (GCR). The direct risks through irreversible extreme ecosystem impacts are currently unknown. No mechanisms for this have been identified. But extreme ecosystem impacts cannot be confidently ruled out given the nature of the Earth systems. SAI could have numerous diffuse impacts on critical systems such as agriculture, politics and health. These currently appear modest, but we cannot rule out the possibilities of systemic cascades or synchronous failures. It appears unlikely that SAI would trigger any other calamitous hazards unless it ignites geopolitical conflict between great powers. Instead, SAI's greatest contribution is through latent risk: the ability for termination shock to significantly worsen any other GCR. For each of these areas the evidence base is significantly underdeveloped.

Is SAI worse than the initial problem of climate change? The question for now is largely unanswerable and lies outside the scope of our analysis. This chapter represents a first step in understanding the multitude of risks of SAI. But critical gaps in understanding of both high-end warming and SAI remain. The climate comparison also depends on specific details, such as level of warming, state of politics, and availability of alternatives to SAI (such as rapid large-scale carbon dioxide removal). SAI is also deeply dependent on governance and the level of use. A constrained use of SAI with coherent, coordinated governance would most likely be benign and beneficial. Yet it is in a scenario of extreme warming, political fragmentation, and a search for an escape clause that SAI use appears most likely. Such thick and uncoordinated use of SAI is unwise and an inappropriate precautionary alternative. We would face a planetary Sword of Damocles.

Notes and References

1 The Royal Society. *Geoengineering the Climate: Science, Governance and Uncertainty*. The Royal Society Publishing (2009). https://doi.org/10.1108/02580541011009815

2 Atkinson, A. *Impact Earth: Asteroids, Comets and Meteors — The Growing Threat*. Virgin (1999); Global Challenges Foundation, *Global Catastrophic Risks Report 2016* (2016).

3 Centeno, Miguel A. et al. 'The emergence of global systemic risk', *Annual Review of Sociology, 41* (2015): 65–85. https://doi.org/10.1146/annurev-soc-073014-112317

4 Parker, Andy and Peter J. Irvine. 'The risk of termination shock from solar geoengineering', *Earth's Future* (March 2018): 1–12. https://doi.org/10.1002/2017EF000735

5 Global Challenges Foundation (2016); Bostrom, Nick and Milan M. Ćirković. *Global Catastrophic Risks* (2008): 554.

6 A single algorithmic system that can perform tasks at a level similar to humans across a broad range of cognitive domains.

7 Avin, Shahar et al. 'Classifying Global Catastrophic Risks', *Futures* (2018). https://doi.org/10.1016/j.futures.2018.02.001; Liu, Hin Yan, Kristian Cedervall Lauta and Matthijs Michiel Maas. 'Governing boring apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research', *Futures* (2018). https://doi.org/10.1016/j.futures.2018.04.009; Baum, Seth D. and Anthony Barrett. 'Towards an integrated assessment of Global Catastrophic Risk', *Catastrophic and Existential Risk: Proceedings of the First Colloquium* (2019): 41–62.

8 IPCC. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*, ed. Christopher B. Field et al. Cambridge University Press (2012). https://doi.org/10.1017/CBO9781139177245; Currie, Adrian and Seán Ó hÉigeartaigh. 'Working together to face humanity's greatest threats: Introduction to the future of research on catastrophic and existential risk', *Futures, 102* (September 2018): 1–5. https://doi.org/10.1016/j.futures.2018.07.003; Liu, Lauta and Maas (2018); Avin et al. (2018).

9 Betz, Gregor. 'Accounting for possibilities in decision making', in *The Argumentative Turn in Policy Analysis* (10th ed.), ed. Sven Ove Hansson and Hirsch Gertrude Hadorn. Springer (2016), pp. 135–69.

10 Wagner, G. and Martin L. Weitzman. *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton University Press (2015); Kunreuther, Howard et al. 'Risk management and climate change', *Nature Climate Change, 3*(5) (2013): 447–50. https://doi.org/10.1038/nclimate1740; Ord, Toby, Rafaela Hillerbrand and Anders Sandberg. 'Probing the improbable: Methodological challenges for risks with low probabilities and high stakes', *Journal of Risk Research, 13*(2) (March 2010): 191–205. https://doi.org/10.1080/13669870903126267

11 Homer-Dixon, Thomas et al. 'Synchronous failure: The emerging causal architecture of global crisis', *Ecology and Society, 20*(3) (2015): 1–16.

12 Helbing, Dirk. 'Globally networked risks and how to respond', *Nature* (2013). https://doi.org/10.1038/nature12047; Homer-Dixon, Thomas. *The Upside of Down: Catastrophe, Creativity, and the Renewal of Civilization*. Island Press (2008).

13 Kemp, Luke et al. 'Climate endgame: Exploring catastrophic climate change scenarios', *Proceedings of the National Academy of Sciences, 119*(34) (23 August 2022). https://doi.org/10.1073/pnas.2108146119

14 McLaren, Duncan. 'Mitigation deterrence and the "moral hazard" of solar radiation management', *Earth's Future, 4*(12) (2016). https://doi.org/10.1002/2016EF000445

15 McLaren, Duncan P. 'Whose climate and whose ethics? Conceptions of justice in solar geoengineering modelling', *Energy Research and Social Science* (2018). https://doi.org/10.1016/j.erss.2018.05.021

16 Jebari, Joseph et al. 'From moral hazard to risk-response feedback', *Climate Risk Management, 33* (2021): 100324. https://doi.org/10.1016/j.crm.2021.100324

17 McKinnon, Catriona. 'Sleepwalking into lock-in? Avoiding wrongs to future people in the governance of solar radiation management research', *Environmental Politics* (2018). https://doi.org/10.1080/09644016.2018.1450344

18 This section studies the potential *direct* catastrophic impact of SAI, not the impacts of termination shock (which are almost guaranteed to be catastrophic if used to mask high levels of warming and if SAI undergoes a indefinite suspension. Trisos, Christopher H. et al. 'Potentially dangerous consequences for biodiversity of solar geoengineering implementation and termination', *Nature Ecology and Evolution, 2*(3) (2018). https://doi.org/10.1038/s41559-017-0431-0. By "directly", we refer to clear causal relationships with less than two degrees of separation.

19 Reynolds, Jesse L., Andy Parker and Peter Irvine. 'Five solar geoengineering tropes that have outstayed their welcome Earth's future', *Earth's Future, 4* (2016): 562–68. https://doi.org/10.1002/eft2.158; Tilmes, Simone et al. 'The hydrological impact of geoengineering in the Geoengineering Model Intercomparison Project (GeoMIP)', *Journal of Geophysical Research: Atmospheres, 118*(19) (October 2013): 11–36. https://doi.org/10.1002/jgrd.50868

20 Jiang, Jiu et al. 'Stratospheric sulfate aerosol geoengineering could alter the high-latitude seasonal cycle', *Geophysical Research Letters, 46*(23) (December 2019): 14153–63. https://doi.org/10.1029/2019GL085758

21 Dagon, Katherine and Daniel P. Schrag. 'Exploring the effects of solar radiation management on water cycling in a coupled land–atmosphere model*', *Journal of Climate, 29*(7) (April 2016): 2635–50. https://doi.org/10.1175/JCLI-D-15-0472.1; Dagon, Katherine and Daniel P. Schrag. 'Quantifying the effects of solar geoengineering on vegetation', *Climatic Change* (2019). https://doi.org/10.1007/s10584-019-02387-9

22 Zarnetske, Phoebe L. et al. 'Potential ecological impacts of climate intervention by reflecting sunlight to cool Earth', *Proceedings of the National Academy of Sciences, 118*(15) (2021). https://doi.org/10.1073/pnas.1921854118

23 Zarnetske et al. (2021).

24 Trisos et al. (2018); Zarnetske et al. (2021).

25 Lee, Walker Raymond et al. 'High-latitude stratospheric aerosol geoengineering can be more effective if injection is limited to spring', *Geophysical Research Letters* (n.d.). https://doi.org/10.1002/essoar.10505988.1

26 MacMartin, Douglas G. et al. 'The climate response to stratospheric aerosol geoengineering can be tailored using multiple injection locations', *Journal of Geophysical Research: Atmospheres, 122*(23) (December 2017): 12, 512–74, 590. https://doi.org/10.1002/2017JD026868

27 MacMartin, Douglas G. et al. 'Geoengineering with stratospheric aerosols: what do we not know after a decade of research?', *Earth's Future* (2016). https://doi.org/10.1002/2016EF000418; McCormack, Caitlin G. et al. 'Key impacts of climate engineering on biodiversity and ecosystems, with priorities for future research',

Journal of Integrative Environmental Sciences (2016). https://doi.org/10.1080/1943815X.2016.1159578; Irvine, Peter J. et al. 'Towards a comprehensive climate impacts assessment of solar geoengineering', *Earth's Future, 5*(1) (January 2017): 93–106. https://doi.org/10.1002/2016EF000389; Irvine, Peter J. et al. 'An overview of the Earth system science of solar geoengineering', *Wiley Interdisciplinary Reviews: Climate Change* (2016). https://doi.org/10.1002/wcc.423; Schäfer, Stefan et al. *The European Transdisciplinary Assessment of Climate Engineering* (2015).

28 Brysse, Keynyn et al. 'Climate change prediction: Erring on the side of least drama?', *Global Environmental Change, 23*(1) (2013): 327–37. https://doi.org/10.1016/j.gloenvcha.2012.10.008

29 Jehn, Florian Ulrich et al. 'Betting on the best case: Higher end warming is underrepresented in research', *Environmental Research Letters, 16*(8) (1 August 2021): 084036. https://doi.org/10.1088/1748-9326/ac13ef; Geden, Oliver. 'Policy: Climate advisers must maintain integrity', *Nature, 521*(7550) (May 2015): 27–28. https://doi.org/10.1038/521027a

30 Low, Sean and Matthias Honegger. 'A precautionary assessment of systemic projections and promises from sunlight reflection and carbon removal modeling', *Risk Analysis* (n.d.). https://doi.org/10.1111/risa.13565

31 Irvine, Peter et al. 'Halving warming with idealized solar geoengineering moderates key climate hazards', *Nature Climate Change* (March 2019). https://doi.org/10.1038/s41558-019-0398-8

32 Lee et al. (n.d.).

33 Abatayo, Anna Lou et al. 'Solar geoengineering may lead to excessive cooling and high strategic uncertainty', *Proceedings of the National Academy of Sciences, 117*(24) (2020): 13393–98. https://doi.org/10.1073/pnas.1916637117

34 Kravitz, Ben et al. 'Sulfuric acid deposition from stratospheric geoengineering with sulfate aerosols', *Journal of Geophysical Research: Atmospheres, 114*(D14) (July 2009). https://doi.org/10.1029/2009JD011918; Visioni, Daniele et al. 'What goes up must come down: Impacts of deposition in a sulfate geoengineering scenario', *Environmental Research Letters* (2020). https://doi.org/10.1088/1748-9326/ab94eb

35 Commonly proposed coolant agents could be benign or even increase ozone thickness. Irvine et al. (2016); Pitari, Giovanni et al. 'Stratospheric ozone response to sulfate geoengineering: results from the Geoengineering Model Intercomparison Project (GeoMIP)', *Journal of Geophysical Research: Atmospheres, 119*(5) (March 2014): 2629–53. https://doi.org/10.1002/2013JD020566 (though this is a relatively recent shift).

36 Heckendorn, P. et al. 'The impact of geoengineering aerosols on stratospheric temperature and ozone', *Environmental Research Letters* (2009). https://doi.org/10.1088/1748-9326/4/4/045108; Keith, David W. et al. 'Stratospheric solar geoengineering without ozone loss', *Proceedings of the National Academy of Sciences, 113*(52) (2016): 14910–14. https://doi.org/10.1073/pnas.1615572113

37 Halstead, John. 'Stratospheric aerosol injection research and existential risk', *Futures, 102* (March 2018): 63–77. https://doi.org/10.1016/j.futures.2018.03.004

38 Elsawah, Sondoss et al. 'Eight grand challenges in socio-environmental systems modeling', *Socio-Environmental Systems Modelling, 2* (January 2020): 16226. https://doi.org/10.18174/sesmo.2020a16226

39 A "free-driver" problem (or "forced rider") is where a single or limited number of actors' actions create negative consequences that the rest of the system must endure. In this case, a limited number of actors could implement SAI, and the rest of the world's economic and political systems would have to deal with the consequences. This can be seen as the "opposite" to a conventional "free rider" problem where benefits flow onto non-cooperative third parties,

40 Weitzman, Martin L. 'A voting architecture for the governance of free-driver externalities, with application to geoengineering', *Scandinavian Journal of Economics, 117*(4) (2015): 1049–68. https://doi.org/10.1111/sjoe.12120

41 Seto, Karen C. et al. 'Carbon lock-in: Types, causes, and policy implications', *Annual Review of Environment and Resources, 41*(1) (October 2016): 425–52. https://doi.org/10.1146/annurev-environ-110615-085934

42 Williams, John W., Alejandro Ordonez and Jens-Christian Svenning. 'A unifying framework for studying and managing climate-driven rates of ecological change', *Nature Ecology & Evolution, 5*(1) (2021): 17–26. https://doi.org/10.1038/s41559-020-01344-5; Williams, John W. and Stephen T. Jackson. 'Novel climates, no-analog communities, and ecological surprises', *Frontiers in Ecology and the Environment, 5*(9) (November 2007): 475–82. https://doi.org/10.1890/070037

43 McKinnon (2018).

44 Baum, Seth, Timothy M. Maher Jr. and Jacob Haqq-Misra. 'Double catastrophe: Intermittent stratospheric geoengineering induced by societal collapse', *Environment, Systems and Decisions, 33*(1) (2013): 1–21.

45 Instead of Global Catastrophic Risks, this section focuses on Global Catastrophic Hazards. Risks include vulnerability and exposure. Instead, this section focuses on the interactions between different specific hazards and SAI.

46 Phillips, Carly A. et al. 'Compound climate risks in the COVID-19 pandemic', *Nature Climate Change, 10*(7) (2020): 586–88. https://doi.org/10.1038/s41558-020-0804-2

47 Baum, Maher Jr. and Haqq-Misra (2013).

48 Laakso, A. et al. 'Radiative and climate impacts of a large volcanic eruption during stratospheric sulfur geoengineering', *Atmospheric Chemistry and Physics, 16*(1) (January 2016): 305–23. https://doi.org/10.5194/acp-16-305-2016

49 MacMartin, Douglas G. et al. 'Technical characteristics of a solar geoengineering deployment and implications for governance', *Climate Policy, 19*(10) (November 2019): 1325–39. https://doi.org/10.1080/14693062.2019.1668347

50 MacMartin et al. (2019).

51 MacMartin et al. (2019).

52 Mani, Lara, Asaf Tzachor and Paul Cole. 'Global Catastrophic Risk from lower magnitude volcanic eruptions', *Nature Communications, 12*(1) (6 December 2021): 4756. https://doi.org/10.1038/s41467-021-25021-8

53 Green, James L. and Scott Boardsen. 'Duration and extent of the great auroral storm of 1859', *Advances in Space Research, 38*(2) (2006): 130–35. https://doi.org/10.1016/j.asr.2005.08.054

54 Eastwood, J. P. et al. 'The economic impact of space weather: where do we stand?', *Risk Analysis, 37*(2) (February 2017): 206–18. https://doi.org/10.1111/risa.12765; Ritter, Scott et al. 'International legal and ethical issues of a future carrington event: Existing frameworks, shortcomings, and recommendations', *New Space, 8*(1) (2020):

23–30. https://doi.org/10.1089/space.2019.0026; Loper, Robert D. ‘Carrington-class events as a great filter for electronic civilizations in the drake equation’, *Publications of the Astronomical Society of the Pacific, 131*(998) (April 2019): 044202. https://doi.org/10.1088/1538-3873/ab028e

55 There is unresolved discussion as to whether extreme space weather events follow a power law or lognormal distribution. Kataoka, Ryuho. ‘Extreme geomagnetic activities: A statistical study’, *Earth, Planets and Space, 72*(1) (2020): 124. https://doi.org/10.1186/s40623-020-01261-8; Riley, Pete and Jeffrey J. Love. ‘Extreme geomagnetic storms: Probabilistic forecasts and their uncertainties’, *Space Weather, 15*(1) (January 2017): 53–64. https://doi.org/10.1002/2016SW001470. The estimated probabilities of an extreme space weather event in the next decade or so range from 0.46 % to 1.88 %. Moriña, David et al. ‘Probability estimation of a Carrington-like geomagnetic storm’, *Scientific Reports, 9*(1) (2019): 2393. https://doi.org/10.1038/s41598-019-38918-8; 4–6 %, Kataoka, Ryuho. ‘Probability of occurrence of extreme magnetic storms’, *Space Weather, 11*(5) (May 2013): 214–18. https://doi.org/10.1002/swe.20044, to roughly 20.3 %. Riley and Love (2017). This is all to say that while there are efforts to better understand the probability of extreme space weather events, there is little agreement in what the precise probability is. Nonetheless, even the lowest estimates are not negligible. In the face of such uncertainty, we take the lead of other policy analyses in the area. Royal Academy of Engineering. *Extreme Space Weather: Impacts on Engineered Systems and Infrastructure* (2013); National Science and Technology Council. *National Space Weather Strategy and Action Plan* (2019) and characterise these events as essentially random.

56 Royal Academy of Engineering (2019).

57 Jones, J. B. L. et al. ‘Space weather and commercial airlines’, *Advances in Space Research, 36*(12) (2005): 2258–67. https://doi.org/10.1016/j.asr.2004.04.017

58 Goodman, John M. ‘Operational communication systems and relationships to the ionosphere and space weather’, *Advances in Space Research, 36*(12) (2005): 2241–52. https://doi.org/10.1016/j.asr.2003.05.063; Royal Academy of Engineering (2019).

59 Jones et al. (2005).

60 Alvarez, Luis E., Sebastian D. Eastham and Steven R. H. Barrett. ‘Radiation dose to the global flying population’, *Journal of Radiological Protection, 36*(1) (March 2016): 93–103. https://doi.org/10.1088/0952-4746/36/1/93; Dyer, C. S. et al. ‘Solar particle enhancements of single-event effect rates at aircraft altitudes’, *IEEE Transactions on Nuclear Science, 50*(6) (December 2003): 2038–45. https://doi.org/10.1109/TNS.2003.821375; Jones et al. (2005).

61 Jones et al. (2005).

62 Royal Academy of Engineering (2019); National Science and Technology Council (2019).

63 Liu, Ying D. et al. ‘Observations of an extreme storm in interplanetary space caused by successive coronal mass ejections’, *Nature Communications, 5*(1) (2014): 3481. https://doi.org/10.1038/ncomms4481; Pulkkinen, Antti et al. ‘Regional-scale high-latitude extreme geoelectric fields pertaining to geomagnetically induced currents’, *Earth, Planets and Space, 67*(1) (2015): 93. https://doi.org/10.1186/s40623-015-0255-6; Eastwood et al. (2017).

64 Eroshenko, E. A. et al. ‘Effects of strong geomagnetic storms on northern railways in Russia’, *Advances in Space Research, 46*(9) (2010): 1102–10. https://doi.org/10.1016/j.asr.2010.05.017; Wik, M. et al. ‘Space weather events in july 1982 and october 2003 and

the effects of geomagnetically induced currents on Swedish technical systems', *Annales Geophysicae, 27*(4) (2009): 1775–87. https://doi.org/10.5194/angeo-27-1775-2009; Ptitsyna, N. G. et al. 'Geomagnetic effects on mid-latitude railways: A statistical study of anomalies in the operation of signaling and train control equipment on the East-Siberian Railway', *Advances in Space Research, 42*(9) (2008): 1510–14. https://doi.org/10.1016/j.asr.2007.10.015

65 Juusola, L. et al. 'High-latitude ionospheric equivalent currents during strong space storms: Regional perspective', *Space Weather — The International Journal Of Research And Applications, 13*(1) (January 2015): 49–60. https://doi.org/10.1002/2014SW001139; Wang, Kai-rang, L. I. U. Lian-guang and L. I. Yan. 'Preliminary analysis on the interplanetary cause of geomagnetically induced current and its effect on power systems', *Chinese Astronomy and Astrophysics, 39*(1) (2015): 78–88. https://doi.org/10.1016/j.chinastron.2015.01.003; Matandirotya, Electdom, Pierre J. Cilliers and Robert R. Van Zyl. 'Modeling geomagnetically induced currents in the South African power transmission network using the finite element method', *Space Weather, 13*(3) (March 2015): 185–95. https://doi.org/10.1002/2014SW001135; Royal Academy of Engineering (2019).

66 Odenwald, Sten, James Green and William Taylor. 'Forecasting the impact of an 1859-calibre superstorm on satellite resources', *Advances in Space Research, 38*(2) (2006): 280–97. https://doi.org/10.1016/j.asr.2005.10.046

67 Loper (2019).

68 Loper (2019).

69 MacMartin et al. (2019).

70 National Science and Technology Council (2019).

71 Barrett, Anthony M., Seth D. Baum and Kelly Hostetler. 'Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia', *Science & Global Security, 21*(2) (May 2013): 106–33. https://doi.org/10.1080/08929882.2013.798984; Baum, Seth, Robert de Neufville and Anthony Barrett. 'A model for the probability of nuclear war', *SSRN Electronic Journal* (2018). https://doi.org/10.2139/ssrn.3137081

72 Baum, de Neufville and Barrett (2018).

73 Baum, Maher Jr. and Haqq-Misra (2013).

74 Edwards, Paul N. 'Entangled histories: Climate science and nuclear weapons research', *Bulletin of the Atomic Scientists, 68*(4) (July 2012): 28–40. https://doi.org/10.1177/0096340212451574

75 Jinnah, Sikina. 'Building a governance foundation for solar geoengineering deployment', in *Governance of the Deployment of Solar Geoengineering*, ed. Robert N. Stavins and Robert Stowe. Harvard Project on Climate Agreements (2019), pp. 143–48; Horton, J. B. and J. L. Reynolds. 'The international politics of climate engineering: a review and prospectus for international relations', *International Studies Review, 18*(3) (2016): 438–61. https://doi.org/10.1093/isr/viv013; Chhetri, Netra et al. *Governing Solar Radiation Management* (2018). https://doi.org/10.17606/M6SM17

76 McLaren, Duncan and Olaf Corry. 'Clash of geofutures and the remaking of planetary order: Faultlines underlying conflicts over geoengineering governance', *Global Policy, 12*(S1) (April 2021): 20–33. https://doi.org/10.1111/1758-5899.12863

77 Doan, Xuan Vinh and Duncan Shaw. 'Resource allocation when planning for simultaneous disasters', *European Journal of Operational Research, 274*(2) (2019):

687–709. https://doi.org/10.1016/j.ejor.2018.10.015; Platt, Rutherford H. *Disasters and Democracy: The Politics of Extreme Natural Events*. Island Press (1999); Cohen, Charles and Eric D. Werker. 'The political economy of "natural" disasters', *Journal of Conflict Resolution, 52*(6) (December 2008): 795–819. https://doi.org/10.1177/0022002708322157

78 Baum, Maher Jr. and Haqq-Misra (2013).

79 Carlson, Colin J. et al. 'Solar geoengineering could redistribute malaria risk in developing countries', *MedRxiv* (2020). https://doi.org/10.1101/2020.10.21.20217257; Carlson, Colin J. and Christopher H. Trisos. 'Climate engineering needs a clean bill of health', *Nature Climate Change, 8*(10) (2018): 843–45. https://doi.org/10.1038/s41558-018-0294-7

80 Christophersen, Olav Albert and Anna Haug. 'Why is the world so poorly prepared for a pandemic of hypervirulent avian influenza?', *Microbial Ecology in Health and Disease, 18*(3–4) (January 2006): 113–32. https://doi.org/10.1080/08910600600866544; Morens, David M. et al. 'The origin of COVID-19 and why it matters', *The American Journal of Tropical Medicine and Hygiene, 103*(3) (September 2020): 955–59. https://doi.org/10.4269/ajtmh.20-0849; Cheng, Vincent C. C. et al. 'Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection', *Clinical Microbiology Reviews* (2007). https://doi.org/10.1128/CMR.00023-07; Quammen, David. *Spillover: Animal Infections and the Next Human Pandemic*. W. W. Norton & Company (2012); Plowright, Raina K. et al. 'Pathways to Zoonotic spillover', *Nature Reviews Microbiology, 15*(8) (2017): 502–10. https://doi.org/10.1038/nrmicro.2017.45; Johnson, Christine K. et al. 'Global shifts in mammalian population trends reveal key predictors of virus spillover risk', *Proceedings of the Royal Society B: Biological Sciences, 287*(1924) (April 2020): 20192736. https://doi.org/10.1098/rspb.2019.2736; Borremans, Benny et al. 'Cross-species pathogen spillover across ecosystem boundaries: Mechanisms and theory', *Philosophical Transactions of the Royal Society B: Biological Sciences, 374*(1782) (September 2019): 20180344. https://doi.org/10.1098/rstb.2018.0344

81 Helbing (2013); Homer-Dixon (2008); Haldane, Andrew G. and Robert M. May. 'Systemic risk in banking ecosystems', *Nature, 469*(7330) (2011): 351–55. https://doi.org/10.1038/nature09659

82 Helbing (2013).

83 Russon, Mary-Ann. 'The cost of the Suez Canal blockage', *BBC News* (2021).

84 We exclude other areas such as telecommunications and finance due to a lack of published literature.

85 Kortetmäki, Teea and Markku Oksanen. 'Food systems and climate engineering : A plate full of risks or promises?', *Climate Justice and Geoengineering*. Rowman & Littlefield International (2016), pp. 121–35; Pamplany, Augustine, Bert Gordijn and Patrick Brereton. 'The ethics of geoengineering: A literature review', *Science and Engineering Ethics, 26*(6) (2020): 3069–3119. https://doi.org/10.1007/s11948-020-00258-6; Svoboda, Toby et al. 'The potential for climate engineering with stratospheric sulfate aerosol injections to reduce climate injustice', *Journal of Global Ethics, 14*(3) (2018). https://doi.org/10.1080/17449626.2018.1552180; Irvine et al. (2017).

86 Xia, Lilia et al. 'Solar radiation management impacts on agriculture in China: A case study in the Geoengineering Model Intercomparison Project (GeoMIP)', *Journal of Geophysical Research: Atmospheres, 119*(14) (July 2014): 8695–8711. https://doi.org/10.1002/2013JD020630

[87] Pongratz, J. et al. 'Crop yields in a geoengineered climate', *Nature Climate Change, 2*(2) (2012): 101–5. https://doi.org/10.1038/nclimate1373

[88] Yang, Huiyi et al. 'Potential negative consequences of geoengineering on crop production: A study of Indian groundnut', *Geophysical Research Letters, 43*(22) (November 2016): 11,711–86,795. https://doi.org/10.1002/2016GL071209

[89] Proctor, Jonathan et al. 'Estimating global agricultural effects of geoengineering using volcanic eruptions', *Nature, 560*(7719) (2018): 480–83. https://doi.org/10.1038/s41586-018-0417-3

[90] Richards, C. E., R. C. Lupton and J. M. Allwood. 'Re-framing the threat of global warming: An empirical causal loop diagram of climate change, food insecurity and societal collapse', *Climatic Change, 164*(3) (2021): 49. https://doi.org/10.1007/s10584-021-02957-w; Natalini, Davide, Giangiacomo Bravo, and Aled Wynne Jones. 'Global food security and food riots — An agent-based modelling approach', *Food Security, 11* (2017): 1153–73. https://doi.org/10.1007/s12571-017-0693-z; Natalini, Davide, Aled Wynne Jones and Giangiacomo Bravo. 'Quantitative assessment of political fragility indices and food prices as indicators of food riots in countries', *Sustainability, 7*(4) (2015): 4360–85. https://doi.org/10.3390/su7044360

[91] Kortetmäki and Oksanen (2016).

[92] Proctor et al. (2018).

[93] Xu, Chi et al. 'Future of the human climate niche', *Proceedings of the National Academy of Sciences* (2020): 1–6. https://doi.org/10.1073/pnas.1910114117; Burke, Marshall, Solomon M. Hsiang and Edward Miguel. 'Global non-linear effect of temperature on economic production', *Nature, 527*(7577) (2015): 235–39. https://doi.org/10.1038/nature15725

[94] McKinnon, Catriona. 'The Panglossian politics of the geoclique', *Critical Review of International Social and Political Philosophy, 23*(5) (2020): 584–99. https://doi.org/10.1080/13698230.2020.1694216; Horton and Reynolds (2016).

[95] Corry, Olaf. 'The international politics of geoengineering: The feasibility of Plan B for tackling climate change', *Security Dialogue, 48*(4) (2017): 297–315. https://doi.org/10.1177/0967010617704142; Keith, David. *A Case for Climate Engineering*. MIT Press (2013).

[96] MacMartin et al. (2019).

[97] Horton and Reynolds (2016); Fleming, James Rodger. 'The climate engineers', *The Wilson Quarterly, 31*(2) (2007): 46–60. https://doi.org/10.2307/40262106; Lin, Albert C. 'The missing pieces of geoengineering research governance', *Minnesota Law Review, 100* (2016): 2509–76; Olson, R. 'Geoengineering for decision makers', *Science and Technology Innovation Program* (2011); Halstead (2018).

[98] BBC, 'US fuel pipeline "paid hackers $5m in ransom"', *BBC News* (2021).

[99] Rolnick, David et al. 'Tackling climate change with machine learning', *ArXiv* (2019); de Witt, Christian Schroeder and Thomas Hornigold. 'Stratospheric aerosol injection as a deep reinforcement learning problem', *ArXiv* (2019).

[100] MacMartin et al. (2019).

[101] Slay, Jill and Michael Miller. *Lessons Learned from the Maroochy Water Breach BT — Critical Infrastructure Protection*, ed. Eric Goetz and Sujeet Shenoi. Springer US (2008), pp. 73–82.

[102] Wells, Linton. *Thoughts for the 2001 Quadrennial Defense Review* (2001).

103 Carlson and Trisos (2018).

104 Carlson et al. (2020); Carlson and Trisos (2018).

105 Effiong, Utibe and Richard L. Neitzel. 'Assessing the direct occupational and public health impacts of solar radiation management with stratospheric aerosols', *Environmental Health, 15*(1) (December 2016): 7. https://doi.org/10.1186/s12940-016-0089-0; Eastham, Sebastian D. et al. 'Quantifying the impact of sulfate geoengineering on mortality from air quality and UV-B exposure', *Atmospheric Environment, 187* (2018): 424–34. https://doi.org/10.1016/j.atmosenv.2018.05.047; Carlson et al. (2020); Carlson and Trisos (2018).

106 Effiong and Neitzel (2016).

107 Eastham et al. (2018).

108 Carlson and Trisos (2018).

109 Cotton-Barratt, Owen, Max Daniel and Anders Sandberg. 'Defence in depth against human extinction: Prevention, response, resilience, and why they all matter', *Global Policy, 11*(3) (1 May 2020): 271–82. https://doi.org/10.1111/1758-5899.12786

110 Kosugi, Takanobu. 'Fail-safe solar radiation management geoengineering', *Mitigation and Adaptation Strategies for Global Change, 18*(8) (2013): 1141–66. https://doi.org/10.1007/s11027-012-9414-2

111 Parker and Irvine (2018).

112 Lynas, Mark. *Our Final Warning: Six Degrees of Climate Emergency*. Harper Collins (2020).

113 Parker and Irvine (2018); Reynolds, Parker and Irvine (2016).

114 Ott, Konrad K. 'On the political economy of solar radiation management strategies to combat impacts of climate', *Frontiers in Environmental Science, 6* (June 2018): 1–13. https://doi.org/10.3389/fenvs.2018.00043

115 Moriña et al. (2019).

116 Riley and Love (2017).

117 Barrett, Baum and Hostetler (2013).

118 Reynolds, Parker and Irvine (2016).

119 Parker and Irvine (2018).

120 Castillo, Juan Camilo et al. 'Market design to accelerate COVID-19 vaccine supply', *Science, 371*(6534) (12 March 2021): 1107–9. https://doi.org/10.1126/science.abg0889

121 Homer-Dixon et al. (2015).

122 Parker and Irvine (2018); McKinnon (2020); Halstead (2018).

123 Rabitz, Florian. 'Governing the termination problem in solar radiation management', *Environmental Politics, 28*(3) (2019): 502–22. https://doi.org/10.1080/09644016.2018.1519879; Parker and Irvine (2018).

124 Maas, Matthijs M. 'How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons', *Contemporary Security Policy, 40*(3) (July 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464; Perrow, Charles. *Normal Accidents: Living With High Risk Technologies — Updated Edition*. Princeton University Press (1999).

125 Lee, W. et al. 'Expanding the design space of stratospheric aerosol geoengineering to include precipitation-based objectives and explore trade-offs', *Earth System Dynamics, 11*(4) (2020): 1051–72. https://doi.org/10.5194/esd-11-1051-2020; Zarnetske et al. (2021).

126 Kravitz, Ben et al. 'Geoengineering as a design problem', *Earth System Dynamics, 7*(2) (May 2016): 469–97. https://doi.org/10.5194/esd-7-469-2016

127 MacMartin et al. (2019); Kravitz, Ben et al. 'Comparing surface and stratospheric impacts of geoengineering with different SO2 injection strategies', *Journal of Geophysical Research: Atmospheres, 124*(14) (July 2019): 7900–18. https://doi.org/10.1029/2019JD030329; MacMartin et al. (2017); Keith et al. (2016); Pope, F. D. et al. 'Stratospheric aerosol particles and solar-radiation management', *Nature Climate Change, 2*(10) (2012): 713–19. https://doi.org/10.1038/nclimate1528; Visioni, Daniele et al. 'Seasonally modulated stratospheric aerosol geoengineering alters the climate outcomes', *Geophysical Research Letters, 47*(12) (June 2020): e2020GL088337. https://doi.org/10.1029/2020GL088337

128 Bala, G. 'Problems with geoengineering schemes to combat climate change', *Current Science, 96*(1) (2009): 41–48. https://doi.org/10.1006/asle.2002.0099; Krishnamohan, K. S. et al. 'The climatic effects of hygroscopic growth of sulfate aerosols in the stratosphere', *Earth's Future, 8*(2) (February 2020): e2019EF001326. https://doi.org/10.1029/2019EF001326

129 MacMartin et al. (2017).

130 MacMartin et al. (2017).

131 Sun, Weiyi et al. 'Global monsoon response to tropical and arctic stratospheric aerosol injection', *Climate Dynamics, 55*(7) (2020): 2107–21. https://doi.org/10.1007/s00382-020-05371-7

132 Wiertz, Thilo. 'Visions of climate control', *Science, Technology, & Human Values, 41*(3) (May 2016): 438–60. https://doi.org/10.1177/0162243915606524; McLaren (2018).

133 McConnell, Allan. 'Hidden agendas: Shining a light on the dark side of public policy', *Journal of European Public Policy, 25*(12) (December 2018): 1739–58. https://doi.org/10.1080/13501763.2017.1382555

134 Koops, Bert-Jaap. 'The concept of function creep', *Law, Innovation and Technology, 13*(1) (January 2021): 29–56. https://doi.org/10.1080/17579961.2021.1898299

135 There is the ethical question of how research in SAI misuse should be undertaken. Will researching SAI misuse possibilities inadvertently encourage SAI misuse? This is a question beyond the scope of this chapter but is regardless a critical issue that deserves future attention.

136 Talberg, Anita, Sebastian Thomas and John Wiseman. 'A scenario process to inform australian geoengineering policy', *Futures* (June 2018). https://doi.org/10.1016/j.futures.2018.06.003; Preston, Christopher J. 'Ethics and geoengineering: Reviewing the moral issues raised by solar radiation management and carbon dioxide removal', *Wiley Interdisciplinary Reviews: Climate Change, 4*(1) (2013): 23–37. https://doi.org/10.1002/wcc.198

137 Lockley, Andrew and D'Maris Coffman. 'Distinguishing morale hazard from moral hazard in geoengineering', *Environmental Law Review* (2016). https://doi.org/10.1177/1461452916659830; Reynolds, Jesse. 'A critical examination of the climate engineering moral hazard and risk compensation concern', *Anthropocene Review, 2*(2) (2015): 174–91. https://doi.org/10.1177/2053019614554304

138 MacMartin et al. (2019).

139 Rolnick et al. (2019); Schroeder de Witt and Hornigold (2019).

140 Rudin, Cynthia and Joanna Radin. 'Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition', *Harvard Data Science Review, 1*(2) (November 2019). https://doi.org/10.1162/99608f92.5a8a3a3d

141 Christie, Alec P. et al. 'The challenge of biased evidence in conservation', *Conservation Biology* (2020). https://doi.org/10.1111/cobi.13577; Martin, Laura J., Bernd Blossey and Erle Ellis. 'Mapping where ecologists work: Biases in the global distribution of terrestrial ecological observations', *Frontiers in Ecology and the Environment, 10*(4) (2012): 195–201. https://doi.org/10.1890/110154

142 Reynolds, Jesse L. and Gernot Wagner. 'Highly decentralized solar geoengineering', *Environmental Politics* (July 2019): 1–17. https://doi.org/10.1080/09644016.2019.1648169

143 This is a causal loop diagram. A complete line represents a positive polarity (meaning an amplifying feedback, not necessarily positive in a normative sense) and a dotted line denotes a negative polarity (meaning a dampening feedback).

144 Reynolds, Jesse L. 'Solar geoengineering to reduce climate change: A review of governance proposals', *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* (2019). https://doi.org/10.1098/rspa.2019.0255

145 MacMartin et al. (2019).

146 McLaren (2018).

147 McLaren and Corry (2021).

148 There are likely going to be many "in-betweens", but these seem like the major contours of SAI's future.

16. Bioengineering Horizon Scan 2020

Luke Kemp, Laura Adam, Christian R Boehm, Rainer Breitling, Rocco Casagrande, Malcolm Dando, Appolinaire Djikeng, Nicholas G Evans, Richard Hammond, Kelly Hills, Lauren A Holt, Todd Kuiken, Alemka Markotić, Piers Millett, Jonathan A Napier, Cassidy Nelson, Seán S ÓhÉigeartaigh, Anne Osbourn, Megan J Palmer, Nicola J Patron, Edward Perello, Wibool Piyawattanametha, Vanessa Restrepo-Schild, Clarissa Rios-Rojas, Catherine Rhodes, Anna Roessing, Deborah Scott, Philip Shapira, Christopher Simuntala, Robert D J Smith, Lalitha S Sundaram, Eriko Takano, Gwynn Uttmark, Bonnie C Wintle, Nadia B Zahra and William J Sutherland

Highlights:

- Horizon scanning is intended to identify the opportunities and threats associated with technological, regulatory and social change. This chapter presents the second horizon scan for bioengineering conducted by the Centre for the Study of Existential Risk.[1] This was based on inputs from a group of 38 participants from 13 countries across six continents. It identified 20 emerging issues, classified according to the timescale during which they were expected to become most relevant.
- The issues expected to be most relevant within the next five years were: access to biotechnology through outsourcing; crops for changing climates; function-based design in protein engineering; philanthropy shaping bioscience research agendas; and state and international regulation of DNA database use.

https://doi.org/10.11647/OBP.0360.16

- The issues expected to become most relevant in 5–10 years were agricultural gene drives; neuronal probes expanding new sensory capabilities; distributed pharmaceutical development and manufacturing; genetically engineered phage therapy; human genomics converging with computing technologies; microbiome engineering in agriculture; phytoremediation of contaminated soils; production of edible vaccines in plants; and the rise of personalised medicine.
- Finally, the issues expected to become most relevant after 10 or more years were bio-based production of materials; live plant dispensers of chemical signals: the malicious use of advanced neurochemistry; enhancing carbon sequestration; porcine bioengineered replacement organs; and the governance of cognitive enhancement.
- The early identification of such issues is relevant for researchers, policy-makers and the wider public. In addition, participants identified a list of seven underlying themes, highlighting the systemic drivers of change in contemporary bioengineering.

This chapter reproduces one of the largest and most significant pieces of horizon-scanning and expert elicitation work that has been undertaken in the field to-date. It charts some of the most notable emergent themes in bioengineering that may impact the planet and human societies in the coming decades. As a form of foresight, it bears similarities to the other anticipatory and futuring approaches utilised in this volume, in Chapters 11 and 16, whilst its concern with understanding bioengineering risks in relation to political economy, climate change and other areas of converging technological advance draws upon earlier conceptual work such as those explicated in Chapter 3 and 4.

1. Introduction

Bioengineering is expected to have profound impacts on society in the near future as applications increase across multiple areas, while costs and barriers to access fall. The speed of this change and the breadth of the applications make the task of forecasting the impacts of bioengineering both urgent and difficult.[2] In 2017 we published the results of a "horizon scan" that looked at emerging issues in bioengineering.[3] Here we report the results of an updated horizon scan based on a wider range of inputs (38 participants from six continents and 13 countries, compared with 27 participants from the UK and US in the 2017 exercise) and a broader definition of bioengineering.

We followed the same structured "investigate, discuss, estimate and aggregate" (IDEA) protocol for identifying and prioritising issues,[4] with some minor adjustments (see Methods). We tasked our experts with identifying "novel, plausible and high-impact" issues in biological engineering, and they produced a long list of 83 issues. Participants then scored the issues anonymously (with a score out of 1000, reflecting likelihood, impact and novelty), arriving at a short list of issues to be discussed at a workshop. This was coupled with a "yes/no" question to determine whether the issues were novel, based on whether the experts had heard of the issue previously. After deliberation, participants re-scored these issues. The issues identified in the latest horizon scan differ substantially from those identified in 2017. This change likely stems from an increase in the diversity of the participants, improvements in the methods used, a broader definition of bioengineering, and changes in the research landscape since 2017.

Horizon scanning aims to build societal preparedness by systematically identifying upcoming opportunities and threats from technological, regulatory and social change.[5] Horizon scanning with the Delphi technique has a long history. It has been used to identify emerging critical issues in areas as diverse as conservation biology,[6] invasive species in the UK,[7] poverty reduction[8] and biosecurity.[9] Periodic horizon scanning is also undertaken in some areas: in global conservation; for example, these scans have identified issues such as micro-plastics, gene editing for invasive species, and cultivated meat approximately six years before they captured public attention.[10] Horizon-scanning activities related to the Antarctic and Southern Ocean[11] have also directed funding and policy,[12] and helped to provide the basis for research roadmaps.[13]

There have been developments in a number of the issues identified in the 2017 exercise. Human germline genome editing came to prominence in late 2018 when Chinese researcher He Jiankui announced the birth of two girls with CRISPR/Cas9-edited genomes.[14] Military funding of bioengineering projects also remained substantial: for example, projects funded by DARPA included programs to explore the use bioelectronics for tissue repair and regeneration (BETR) and to develop mosquito-repellent skin (ReVector). There have also been breakthroughs in the use of enhanced photosynthesis for agricultural productivity: a 2018 study reported that metabolic engineering strategies increased photosynthetic efficiency by 17%, which resulted in an increase of about 20% in biomass in field conditions.[15] This technology is now being deployed in several crops. Platform technologies to address emerging disease pandemics have taken on particular significance with the advent of the COVID-19 pandemic. Many of the rapidly created vaccine candidates for COVID-19 in clinical and pre-clinical evaluation have been efficiently developed from platforms for non-Coronavirus candidates such as influenza, SARs and Ebola.[16]

In this article we provide a high-level summary of the top 20 issues identified in the bioengineering horizon scan 2020 (while acknowledging that the number of topics covered means that there will be some sacrifice of depth for breadth). We take a broader view of bioengineering than we did in 2017, defining it as the application of ideas, principles and techniques to the engineering of biological systems. This means that we now cover more aspects of bioengineering, as well as issues that contribute to or result from bioengineering advances (such as funding). To avoid giving a false sense of forecasting precision or overemphasising minor differences in scoring, the issues are not ranked, and are instead grouped into issues that are expected to be most relevant within five years, within 5–10 years, and on timescales of longer than 10 years (Table 1). Our intent is to spur further research into these issues and further discussion of their implications by researchers, policy-makers and the wider public.

Table 1: Overview of the bioengineering horizon scan 2020. Summary of the 20 issues identified through the scan; issues are grouped according to likely timeline for realisation.

< 5 Years	**5–10 Years**	**>10 Years**
Access to biotechnology through outsourcing	Agricultural gene drives	Bio-based production of materials
Crops for changing climates	Neuronal probes expanding new sensory capabilities	Live plant dispensers of chemical signals
Function-based design in protein engineering	Distributed pharmaceutical development and manufacturing	Malicious use of advanced neurochemistry
Philanthropy shapes bioscience research agendas	Genetically engineered phage therapy	Enhancing carbon sequestration
State and international regulation of DNA database use	Human genomics converging with computing technologies	Porcine bioengineered replacement organs
	Microbiome engineering in agriculture	The governance of cognitive enhancement
	Phytoremediation of contaminated soils	
	Production of edible vaccines in plants	
	The rise of personalised medicine such as cell therapies	

2. The Issues Most Relevant Within Five Years

2.1 Access to biotechnology through outsourcing

Traditionally, the biotechnology sector has had high barriers to entry, with organisations needing to build extensive physical and knowledge-based assets. New "cloud labs" and services labs are circumventing this model by using technologies such as robotics, automation and the internet to offer widely accessible standardised services with limited need for physical material transfer.[17] This facilitates both broader access and faster development of new products through the sharing of capital and knowledge across projects.[18] It is also helping to empower non-traditional researchers by lowering the threshold for participating in cutting-edge research.

This distributed approach poses a biosecurity gap as research activities are separated from intent: the cloud lab may not seek additional details on an experiment's context, including why it is being performed. There is a lack of appropriate biosecurity guidelines and governance models to handle this.[19] As outsourcing through cloud labs becomes increasingly prevalent in the next five years, these challenges may require the development of new guidelines and business and incentive models for responsible innovation and biosecurity.

2.2 Crops for changing climates

Climate change is predicted to result in more frequent droughts and intensive precipitation events. This will increase soil salinity, elevate average temperatures, and shift the range, abundance and genotypic diversity of pollinators, pests and pathogens. All of these factors are expected to impact crop yields. In response, efforts are intensifying to adapt food production using agro-ecological strategies,[20] as well as the provision of well-adapted crop varieties by genetic engineering and new breeding technologies:[21] drought-tolerant genetically modified (GM) plant varieties have reached the market and more are in development;[22] the capabilities of plant immune receptors have been broadened by protein engineering;[23] and the identification of conserved submergence-activated genes has revealed novel genetic targets for enhancing flood tolerance.[24] Technical progress is still required for success in the field. However, deployment may be hindered by a comparative lack of funding for plant science, as well as lengthy and expensive regulatory regimes in most jurisdictions. New models for public-private co-operation will be needed to advance the translation of basic research through to the field, including business models that are not based on simple economic returns. The effects of novel traits on biodiversity and ecosystems will require further scrutiny before being deployed in a warmer world.

2.3 Function-based design in protein engineering

Despite a growing understanding of the relationship between protein structure and function, efficient design of new proteins with a desired action has remained a laborious process. For example, chimeric antigen receptor (CAR) thymus lymphocyte (T cell) therapies which combine functional protein moieties to activate T cells against malignant tumours have only recently been approved for human use after decades of iteration.[25] The convergence of ongoing developments, including substantial improvements in predicting protein structure from amino acid sequences using machine learning,[26] could overcome previous technical and computational challenges. This indicates a potential revolution in function-based protein design, leading to various useful industrial compounds (such as the development of catalysts for any desired organic reaction) and medical applications (such as the ability to selectively destroy, suppress or stimulate any malfunctioning tissue, which is the key to treating many refractory diseases). However, as this field grows, so will the risk of deliberate misuse. Protein engineering could be used to produce agents that have a higher lethality or specificity than existing agents (including new agents based on novel mechanisms of action). Protein engineering might also simplify the production of toxins currently derived from natural sources.

2.4 Philanthropy shapes bioscience research agendas

Over the past decade, philanthropic funding (including venture philanthropy) of research and innovation has been increasing.[27] This has largely been driven by the increasing concentration of wealth, and erosion of public health and scientific research initiatives within key countries. These investments can provide particular research groups or areas with substantial funding over prolonged periods of time, and they can also support areas of research that are not usually funded by governments. Philanthropic investments can also promote innovation, such as allowing for more exotic approaches not usually funded by governments. However, these investments might also influence the development of biotechnologies in a way that has less of a public mandate than government-funded research and operate without traditional mechanisms for accountability, transparency or oversight often required by federal or state law.[28] Some areas of medical research are already considerably underfunded compared to health needs,[29] and philanthropic investments may exacerbate this discrepancy in the near-term future. Significant investment into a small range of actors could also undermine diversity, particularly at the international level.[30]

2.5 State and international regulation of deoxyribonucleic acid (DNA) database use

Personal genomic sequencing continues to drop in price and increase in accessibility. The inherent inability to truly anonymise such data, coupled with the wealth of information it provides on both individuals and families, distinguishes it from conventional data types such as fingerprints (identifiable but uninformative) or shopping habits.[31] The drop in price and the use of enabling technologies such as cloud storage has enabled wider use of DNA databases by different actors. While the vulnerability of cloud infrastructure is a concern, there is greater potential for misuse by states and law enforcement in the name of security. This has been seen in efforts to target Muslim Uighurs in China via blood samples,[32] and in a consumer genetics database allowing the Federal Bureau of Investigation in the US to compare genetic data from crime scenes to a database of over two million profiles without customer consent.[33] The potential to accrue and analyse vast amounts of genomic information raises concerns over privacy, especially mass surveillance;[34] the potential expansion of state surveillance powers necessitates dialogue and policy intervention domestically and internationally.

3. Issues Most Relevant in 5–10 Years

3.1 Agricultural gene drives

Gene drives were initially proposed for the control of insect vectors for human diseases,[35] but recent work suggest that they could have major economic benefits agricultural sector.[36] However, while there is potential for gene drives to eliminate or suppress pest species, their widespread uptake and use could lead to problems in their application and governance.[37] One concern is that commercial interests will seek to maintain sales of agrochemicals by configuring gene drives to reduce chemical resistance in target pest insects and weeds as opposed to causing sterility in those species. A second concern is that unilateral deployment of gene drives may cause rapid and unintended ecosystem perturbations without proper oversight or recall. There have also been questions around their control and the lack of public consultation (or participation) regarding their release, as well as legal implications if populations are eliminated within, or new gene configurations are carried to, native locations.[38] Efforts are already underway to counter, control and even reverse the undesired effects of genome editing, including DARPA's Safe Genes program.[39] Policy-makers will need to be vigilant to more problematic applications as agricultural gene drives become more prevalent.

3.2 Neuronal probes expanding new sensory capabilities

New research in creating probes that mimic neurons could enable novel medicinal and enhancement applications such as the creation of new sensory capabilities. Traditionally, neuronal probes have both structural and mechanical dissimilarities from their neuron targets, leading to neuro-inflammatory responses. However, it is now possible to fabricate neuron-like electronic probes (with widths similar to those of neurons) and unobtrusively fuse them with live neurons.[40] Potentially, the technology could be used to add new sensory capability via implanting neuronal probe arrays as a visual cortical prosthesis system. However, such biomimetic sensory probes could introduce unintended vulnerabilities, from a risk of malicious attack via the internet to possible mass monitoring of implanted civilians by law enforcement.[41]

3.3 Distributed pharmaceutical development and manufacturing

Outsourcing and increasingly lower barriers to access in bioengineering are allowing for greater localisation and geographical distribution of the manufacturing and development of pharmaceuticals. Bioengineering offers the capacity to create pharmaceutical compounds or their precursors by genetically modifying organisms to produce them. The prospect of non-traditional pharmaceutical manufacture has gained some traction, but with few tangible results. Barriers to distributed pharmaceutical manufacturing becoming broadly adopted include: the scale of production for individual or community use; appropriate safety standards for manufacturing and administration; interfacing with drug approval pathways. Efforts in non-traditional pharma, such as The Open Insulin Project,[42] are rising in profile and will likely continue, whether individual projects are successful or not. This is supported by the Open Pharma movement which seeks to empower innovation through open-access research and development.[43] That itself may shape regulatory frameworks, and may provide new open or distributed models for drug manufacture. However, in the absence of appropriate norms or regulations,[44] it may also lead to the manufacture, at scale, of drugs that are not vetted for safety, or administered under appropriate clinical guidance.[45]

3.4 Genetically engineered phage therapy

The World Health Organization (WHO) recently reported a worrying lack of new antibiotics to address the dangerous trends of rising resistance to existing antibiotics,[46] and antimicrobial resistance has been identified as a potential global catastrophic risk. Phage therapy as a potential alternative to antibiotic treatment has recently seen a renaissance. In particular, the ability to rapidly engineer phage sequences and phage cocktails opens up the prospect of personalised treatments for tackling genetically diverse infections and overcoming problems of antimicrobial resistance.[47] The technical advances observed in the medical application of phage therapy will also have an impact on other uses of phages as delivery systems in biotechnology. Efforts have also been significantly buoyed by development of easier methods for engineering phage to combat the inevitable evolution of phage resistance in bacteria.[48] However, barriers to widespread commercial use persist, including high costs, instability of the medication, the necessity to type the infection (instead of giving a broad-spectrum pill) and immunogenicity. This makes it more likely for phage therapy to be used as a last resort once other treatments have failed.

3.5 Human genomics converging with computing technologies

Human genomics is increasingly incorporating technologies such as blockchain, cloud computing and machine learning. Firms such as

Amazon and Google offer cloud computing-based storage and data analytics services for the petabytes of genetic data stored online, while companies such as Encrypgen and Nebula use blockchain in systems that reward individuals for sharing their genetic data. Artificial Intelligence and machine learning are enabling deep analysis of thousands of molecules with potential to become future drugs,[49] as well as human genomic data.[50] Most recently, deep learning used molecular structure to predict the efficacy of antibiotic candidates.[51] Some uses of these technologies could help address current privacy concerns. This includes the use of blockchain as well as "secret sharing" techniques, in which sensitive information is divided across multiple servers.[52] However, as they are applied to human genomic data in increasingly powerful and connected ways, additional ethical issues will arise. Enlivened and global discussion on how best to handle societal implications will become necessary.[53]

3.6 Microbiome engineering in agriculture

Progress on microbiome engineering and genomic sequencing could allow for beneficial new applications in agriculture, but also risks. Microbiome engineering and the development of synthetic microbiomes offer wide-ranging uses for mammalian health as well as plant and animal productivity, soil health and disease management. A bottom-up approach to microbiome engineering aims to predictably alter microbiome properties and design functions for agricultural and therapeutic applications. Microbiome engineering strategies could provide alternatives to the use of antibiotics for livestock management.[54] These approaches offer the potential for innovative, sustainable pathways for plant disease suppression by engineering the microbiomes indigenous to agricultural soils.[55] Advances in genome sequencing, metagenomics and synthetic biology have already provided a theoretical framework for constructing synthetic microbiomes with novel functionalities. New methods, such as *in situ* mammalian gut microbiome engineering, could help to overcome existing limitations and offer new capabilities for the future.[56] These new methods and advances can support better design of microbiome modulation strategies in mammalian health and agricultural productivity. Yet, the engineering of agricultural microbiomes on a large scale could also create vulnerabilities towards malicious intervention.

3.7 Phytoremediation of contaminated soils

Research in phytoremediation is leading to the creation of engineered plants that could help recuperate contaminated soils, but further field trials are needed along with discussions about their introduction to and implications for the environment. Certain plant species have natural mechanisms that enable both uptake and tolerance of natural and anthropogenic inorganic pollutants. Identifying, expressing and potentially engineering these traits is receiving increased research interest. Preliminary work on transgenic plants in the lab by overexpression of metal ligands, transporters and specific enzymes has led to successful phytoextractions of pollutants including explosives and heavy metals. However, few experiments have been conducted in the field on contaminated soils,[57] where toxicity of various pollutants and the impact of various environmental factors on the plant-microbiome interaction has limited the success of phytoremediation to date. Realising biotechnological phytoremediation will depend on a number of factors: a more robust systemic understanding of plant-microbiome interactions with pollutants;[58] the survivability of these engineered organisms in the environment; understanding and controlling environmental impacts; and robust societal discussion and carefully designed regulatory regimes.

3.8 Production of edible vaccines in plants

Plants offer a scalable low-cost platform for recombinant vaccine production.[59] The introduction of the oral polio vaccine in the 1960s led to huge interest in developing vaccines that can be delivered without the need for injection. Given that plants are widely consumed, they offer an attractive means of vaccine delivery. Plant-expressed antibodies can protect against dental caries. Similarly, expression of norovirus-like particles in transgenic potatoes could raise antibodies against the virus when the material is consumed.[60] Plant-produced vaccines have also been developed for some animal diseases.[61] Oral delivery with minimal processing has the potential to reduce requirements for extensive frameworks for production, purification, sterilisation, packaging and distribution. A major challenge is the need for improvement of the chemical and physical stability of vaccines during transit through the gut in order to ensure efficacy.[62] Also, commercialisation may be difficult under current regulatory regimes.[63] Moreover, if production is scaled up beyond contained greenhouses, this will require the deliberate environmental (field) release of plants engineered to contain vaccines.

3.9 The rise of personalised medicine such as cell therapies

There is an accelerating trend towards the development and approval of personalized therapeutics. These are medical treatments that are tailored towards individuals, accounting for their likely response based on genomic and epigenetic data. In the US in 2018, 42% of all new drug approvals by the Food and Drugs Administration concerned these treatments.[64] However, significant challenges stand in the way of developing and deploying personalized medicine and cell therapies. These includes issues of delivery logistics and cost. The key factor to clinical adoption of personalised medicine is the value recognition by all healthcare stakeholders. Most personalised medicines are genetically guided interventions that address relatively small subsets of patients with rare genetic mutations. The treatment approaches are sometimes costlier due to their increased sophistication and lower demand. Once these barriers are overcome there will be some potential problems that will need to be mitigated via policy. One is ensuring equitable access. Reimbursement from third-party payers such as health insurance companies is also likely to become an issue for targeted treatments.[65] Public health policy must adapt to this new frontier of healthcare while addressing its potentially detrimental effects on equality of healthcare access and treatment.

4. The Issues Most Relevant in 10+ Years

4.1 Bio-based production of materials

Biological engineering and production methods facilitate the transformation of renewable plant feedstocks and microorganisms into substitutes for a wide range of existing and new materials, including plastics and other materials that are produced from fossil fuels.[66] These developments are being driven by increasing government, private and civil society efforts to decarbonise economies. New opportunities may be created for small, bio-based production facilities and clean bio-refineries to be located close to the markets for these materials, potentially replacing much of the petrochemical sector, and there are potential roles for rural areas in growing bio-based feedstocks. While bio-based production promises to be more sustainable than existing methods, attention is still required in addressing specific impacts on feedstocks, energy, water and other environmental and societal factors.[67] This is accompanied by technical barriers in product processing. While some bio-based materials are already on the market, significant private investment and supportive public policy frameworks (including but not limited to carbon pricing, as well as more speculative nitrogen pricing) will be required over the next decade and beyond to accelerate the widespread worldwide transition to these materials.[68]

4.2 Live plant dispensers of chemical signals

Plants emit volatile signals that can activate defence responses in other nearby plants. The concept of using GM plants to deliver these signals has made practical progress in recent years. These genetically modified plants are intended to be helpers that protect surrounding conventional crops that are cultivated for consumption. Field trials have evaluated the potential of transgenic wheat to repel different pests and virus vectors.[69] Despite excellent results in the lab, *in planta* synthesis of the alarm pheromone failed to reduce aphid numbers. Other studies have demonstrated the feasibility of making insect sex pheromones to trap male insects.[70] Further finessing of the pheromone blend may be enabled by synthetic biology. This could open up the possibility of using plants as chemical-producing green factories, or field-based disruptors and dispersers of insect pests. Unlike current GM solutions for protection from insect herbivory, the use of pheromones is a non-lethal and less-persistent intervention, and chemically manufactured pheromones have been in use for many years. Questions remain as to whether the broader adoption of pheromones will simply displace pests to unprotected crops.

4.3 Malicious use of advanced neurochemistry

Agents that could attack the central nervous system were investigated during the Cold War but lack of knowledge only permitted the development of sedating agents. Concerns over such agents and manipulations continues,[71] but could be empowered through advances in neuroscience and other fields. A driving force in these advances is significant government interest and investments, including an investment of almost $1 billion by the US government in the Brain Initiative.[72] Resulting drugs and nootropics offer health benefits, but could also be maliciously used.[73] Governments could use neuro-chemicals to make a populace more subservient. Advanced applications in undeclared biological warfare could include fostering emotional resentment in a targeted population. These drugs could be appealing to governments around the world as a tool for counter-insurgency or non-lethal law enforcement. The use of these new chemicals for law enforcement and in non-traditional conflicts may greatly erode the norms against chemical agent use on the battlefield, threatening the Chemical Weapons Convention in the long term.

4.4 Enhancing carbon sequestration

Metabolic engineering manipulates cells to produce target molecules by optimising endogenous metabolic pathways or by reconstructing these pathways in alternative species. “Next level” metabolic engineering aims to design metabolic networks *de novo*, thus bypassing the bottlenecks and inefficiencies of evolution.[74] Thus far, experimental success is lacking. However, recent research in photosynthesis may be promising, and examples include engineering a novel molecule to realise a designed synthetic photorespiration bypass[75] and developing an optimised carbon dioxide fixation pathway using enzymes from bacteria, archaea, plants and humans.[76] Other methods have included laboratory evolution of a bacterium able to use CO2 for growth.[77] These approaches hold potential for more efficient carbon sequestration and biomass production, as well as for advancing the development of photovoltaics (the production of electricity from light) and light-sustained biomanufacturing. Yet, such developments remain speculative. There are still significant technical challenges to overcome, and a long path to widespread commercial deployment. Moreover, the field will need to engage with its socio-political, ethical, and environmental dimensions.

4.5 Porcine bioengineered replacement organs

Pigs represent a promising candidate species for production of human-compatible replacement organs for xenotransplantation. A recent advance in porcine genome editing using CRISPR/Cas9 addresses one of the key scientific challenges: successful inactivation of porcine endogenous retroviruses, which otherwise pose a risk of cross-species transmission.[78] Such advances hold promise as one technological way to address the global shortage of transplant organs. Over 6500 patients died while on waiting lists in the USA alone in 2017.[79] Several challenges remain, including engineering sufficient immune compatibility in the organs for successful human transplantation, and determining the expected lifespan of the porcine organs in humans. There are differing views over the acceptability of porcine xenotransplantation within major religions, such as Islam and Judaism.[80] Before commercial development, consideration must be given to questions surrounding the ethics of using animals for transplantation, cost and access, and using a technical solution for an essentially social problem that could be addressed through other approaches, such as opt-out organ donation schemes.

4.6 The governance of cognitive enhancement

Cognitive enhancement is already a widely embraced idea throughout society — caffeine is the most widely consumed drug on earth. Novel methods of cognitive enhancement such as nootropics, wakefulness enhancers, or the potential to directly modulate brain function through implants or biotechnology are emerging. Uptake of these is being driven by both a productivity-focused culture, commercial opportunities and increased understanding of neurochemistry. Although some cognitive enhancers require prescriptions, others only have to meet basic safety guidelines and are available to purchase online. While numerous trials have supported the safety of most nootropics and wakefulness enhancers, there are few long-term longitudinal studies.[81] A large section of those who have embraced cognitive enhancement — the "do-it-yourself" experimenters — may also be ignored by the research community. Lax regulation around safety standards for these products and tools has led to calls to tighten regulatory loopholes, and for academic researchers to partner with and include communities in research on cognitive enhancers.[82] Regulatory frameworks are necessary to both minimise risks and gather long-term safety data from end-users, as well as to provide health and safety guidance for international trade of cognitive enhancing drugs and devices.[83]

5. Discussion

5.1 Emergent themes

Seven underlying themes emerged from the workshop discussion: 1) political economy and funding; 2) ethical and regulatory frameworks; 3) climate change; 4) transitioning from lab to field; 5) inequalities; 6) technological convergence; and 7) misuse of technology. None of these were judged precise enough to qualify as horizon-scanning items, although some sub-components were. These themes represent underlying commonalities and drivers across issues.

First, participants expressed concern about the political economy of bioengineering (that is, how political and economic institutions influence bioengineering, including the role of regulation and politics) and, related to this, about funding. These concerns centred around a view that research funded by the military, industry or philanthropy was less accountable than civilian government-funded research and could create real or perceived conflicts of interest.[84]

Second, a recurring theme across several issues was the need for ethics and better regulatory frameworks to manage the problems expected to emerge from technologies on the horizon. This was true for most issues highlighted in the scan, ranging from carbon sequestration to bioengineered replacement organs. This underscores the need for greater engagement between ethicists, social scientists, policy-makers and the cutting-edge of bioengineering.

Third, climate change is likely to be a critical driver of bioengineering in the future. Our list includes an application to both adaptation (crops for changing climates) and negative emissions (sequestration). Others, such as live plant dispensers, could be boosted in relevance as a way to enhance agricultural productivity in the face of detrimental climate impacts. Progress in climate policies will shape the development and demand of bioengineering technologies. Climate change impacts will also create new problems that could be addressed through bioengineering and policy. This includes changes in the range of vector-borne diseases, such as the expansion of tropical infectious diseases.

A fourth theme is that of transitioning from lab to field. The deliberate release of a new bioengineering product into the environment entails risks in both practice and perception. Concerns over the unintended consequences of environmental release have hindered the deployment of GMOs and are now prominent in discussions around gene drives.[85] Such concerns also factored into many of the issues we have identified, most notably edible vaccines and live plant dispensers. Further development of bioengineered products will require appropriate regulation. Additionally, the necessary social, environmental and human health risk assessments need to take place to transition bioengineering from the lab into the wider world.

A fifth theme is the potential for bioengineering to exacerbate existing inequalities in wealth and health. This factored into several issues including the rise of personalised medicine, replacement organs, and the regulation of cognitive enhancement. In contrast, distributed pharmaceutical development and manufacturing was an emerging area fuelled in part by the desire to deliver more equitable, cheap and accessible medicine. Ensuring that the benefits of bioengineering are spread fairly and widely will be a defining feature of future debates. Enhancements also come with risks, especially at the earliest stages. Many of these are expected to be borne by unwilling or uninformed recipients (as in the case of the CRISPR twins) before being marketed to the wealthy. These problems of inequality also highlight the need for horizon-scanning efforts to make efforts to include representatives from more oppressed and marginalised groups.

The sixth theme is that the convergence of different technologies will be crucial in the future development of bioengineering. Many of the issues in this horizon scan are driven by progress in adjacent fields. Both neuronal probes and malicious uses of neurochemistry will be enabled by progress in neuroscience, and the overlap of human genomics with computing technologies brings both opportunities and threats. As automation and measurement, neuroscience, chemistry and artificial intelligence continue, they will shape both what is possible and what is pursued in bioengineering. This poses a challenge for regulators, who may need to think about policy that cuts across bioengineering into other areas, such as cybersecurity. It also highlights a need for continued horizon scanning and foresight exercises to engage a broad range of technological expertise so that key points of intersection and convergence are not overlooked.

Last, our scan highlights ongoing concerns around the misuse of technology by state or non-state actors. Examples included various bioweapons and the misuse of DNA databases.

The 2017 scan noted themes of equality, bioinformatics and regulation, all of which feature prominently in 2020 scan (see Table 2 for a summary of the previous scan). The 2017 exercise discussed the intersection between biotechnology and information and digital technologies. Technological convergence also features in the present scan, but with a broader scope encompassing neuroscience (adding new sensory capabilities) and neurochemistry (malicious uses of advanced neurochemistry) as well as other fields. Both scans featured a strong emphasis on the potential for bioengineering to amplify or alleviate inequalities. In the 2017 scan this included the potential for human genomics to create new "sociogenetic" classes, while differences in healthcare and access to cognitive enhancement were the flagship issues in this 2020 scan. The thematic convergence between the two scans demonstrates that many of the underlying trends in bioengineering include important structural issues involving ethics and regulation. These will likely influence the field for years to come. There were also several differences in themes, including the greater importance of climate change and political economy in the 2020 exercise. This reflects the significant deviation in issues between the two studies.

Table 2: Overview of the bioengineering horizon scan 2017. Summary of the 20 issues identified in 2017; issues are grouped according to likely timeline for realisation.

< 5 Years	**5–10 Years**	**>10 Years**
Artificial photosynthesis and carbon capture for producing biofuels	Regenerative medicine: 3D printing body parts and tissue engineering	New makers disrupt pharmaceutical makers
Enhanced photosynthesis for agricultural productivity	Microbiome-based therapies	Platform technologies to address emerging disease pandemics
New approaches to synthetic gene drives	Producing vaccines and human therapeutics in plants	Challenges to taxonomy-based description and management of biological risk
Human genome editing	Manufacturing illegal drugs using engineered organisms	Shifting ownership models in biotechnology
Accelerating defence agency research in biological engineering	Reassigning codons as genetic firewalls	Securing the critical infrastructure needed to deliver the bioeconomy
	Rise of automated tools for biological design, test and optimisation	
	Biology as information science: impacts on global governance	
	Intersection of information security and bio-automation	
	Effects of the Nagoya Protocol on biological engineering	
	Corporate espionage and biocrime	

Some issues from 2017 also appear in the 2020 exercise in a slightly altered form: concerns about the military use of bioengineering are now more specific (for example, "Malicious use of advanced neurochemistry"), and there are new concerns about the misuse of DNA databases.

Both scans also focussed on different methods for the production of replacement organs. The 2017 exercise identified 3D printing cells on organ-shaped scaffolds, while the 2020 exercise examined the potential for porcine genome editing to allow for xenotransplantation. Finally, both scans assessed the issue of pharmaceutical manufacturing becoming increasingly distributed. The 2017 exercise focused on start-up entrepreneurs and biohacking communities, whereas the 2020 exercise took a broader look at the possibility of decentralisation.

The differences between the scans are likely due to three reasons. First, we used a wider definition of bioengineering which encompassed issues such as biomechanical implants. Two of the issues identified in this scan would not have been covered by the 2017 definition: neuronal probes expanding new sensory capabilities and the governance of cognitive enhancement. Second, half of the participants (19/38) were not involved in the 2017 scan; the new participants were also more geographically diverse (see Methods), and included a higher proportion of social scientists. Third, there have been significant changes in research and the world at large. For example, all the research underpinning the issue of neuronal probes has occurred in the last three years. Similarly, recent research in climate change has highlighted the continued increase in emissions and warming,[86] and that tipping points are more probable than previously expected.[87]

5.2 Limitations and ways forward

While useful, horizon scanning has its limits. Critiques have suggested that the Delphi technique can give unjustified confidence in results that are essentially the subjective judgements of experts.[88] However, in the absence of data, expert elicitation is warranted, and structured approaches such as Delphi and the IDEA protocol have been found to improve group judgement and outperform other forecasting methods, such as prediction markets.[89] While it is difficult to evaluate the efficacy of the Delphi technique due to inconsistencies in its application,[90] those that do exist are promising. A review of a long-term Delphi in predicting developments in the health sector found that results were accurate in 14/18 identified issues.[91] The method continues to show significant utility in both accurately sighting emerging developments and exploring the implications of potential issues on the horizon.

We acknowledge that the issues identified in this horizon scan are ultimately representative of the participants involved. While the 2020 scan is an improvement on previous efforts in terms of diversity, the majority of respondents were still from a developed economy background. The scan did capture a large cross-section of academic subfields in bioengineering, but under-represented industry, communities and policy-makers. Moreover, we achieved a rough gender balance with 21 male participants (55%) and 17 female participants (45%). We intend to make the process increasingly global and diverse under future triennial iterations, and by clearly describing the methods used, have made the process open for uptake by others.

Future pathways for forecasting bioengineering issues are manifold. Further updates of this scan could be paired with systematic reviews of their accuracy and efficacy, as well as deeper dives into the issues that have been identified. Extensions of the horizon-scanning process could include: focusing on specific areas of bioengineering, such as catastrophic risks; incorporating decision-support tools such as fault-trees; examining the development of bioengineering issues in tandem with overlapping technological areas such as artificial intelligence; and producing a policy-focused scan which involves greater engagement with regulators.

6. Methods

Our study made use of the Investigate Discuss Estimate Aggregate (IDEA) protocol. In this process, participants were asked to investigate and submit candidate issues, privately and anonymously score the gathered issues, and discuss their thinking with others. They then provided a second score which was mathematically aggregated. The element of discussion is powerful, as the sharing of information between participants has been shown to improve the accuracy of Delphi-style forecasts.[92] The IDEA protocol has also performed well relative to prediction markets in early studies.[93] Despite being a relatively recent evolution of the Delphi technique, the IDEA protocol has already been successfully applied to a range of areas including natural resource management[94] and assessing pollinator abundance in response to environmental pressures.[95] Aside from seeking a shared understanding of terms and reducing linguistic ambiguity, consensus is not sought during discussion and scores are kept anonymous during both rounds. This is done to avoid undesirable group dynamics and peer pressure distorting individual judgements. Our use of the IDEA Protocol can be split into three phases: i) recruitment and issue gathering; ii) initial scoring; and iii) workshop preparation, deliberation and re-scoring.

6.1 Phase one: Recruitment and issue gathering

Our study drew on a group of 38 participants from six continents. Participants came from countries including the UK, US, Canada, Australia, Germany, Croatia, Thailand, France, Chile, Peru, Switzerland, Malaysia, Zambia and Pakistan. Recruitment was done via a panel of six initial experts (EP, PM, SÓhÉ, CR-R, CR, LS and BW). The panel aimed to ensure a balance across areas such as plant sciences, medicine, bioindustry and biosecurity. They also sought to have a mix of approximately half new participants and half participants from the 2017 exercise scan. Selected bioengineering scholars and practitioners were asked to submit two to five issues each. Our initial request was for issues that were "novel, plausible and high-impact". We asked participants to provide issues that were at a specific level of granularity. As with the previous scan we asked participants not to focus on a general topic, such as "gain of function" research, nor on multiple topics simultaneously. Instead they were guided to focus on one area within a general topic and its implications, such as an emerging regulatory change for GMOs. After duplicates were merged, a long list of 83 issues was generated from the initial submissions. This included 10 merged issues.

6.2 Phase two: Scoring

Participants were asked to vote on the "suitability" of these issues. This involved assigning a score of 0–1000 to each of the issues. Participants were asked to ensure that each score was unique (no identical scores within a given score-sheet). The suitability scores reflected a combination of plausibility, novelty and impact. Novelty was also captured by respondents noting whether they had heard of the issue previously (through a yes/no response). We then calculated the percentage of participants who had heard of each issue. These novelty scores were published alongside all issues in the short list. This was conducted by sending the participants both the long list of issues, along with a template score-sheet and instructions. At this stage participants were reminded that "our aim is to identify plausible, novel bioengineering-related issues with important future implications for society that are not too broad or already well known". They were given approximately three weeks to complete their scoring. Participants were also able to provide comments on the different issues on the voting sheet. These critiques led to a further eight issues being merged into four. Comments were kept to stimulate future discussion. We calculated the z-scores for each participant's issues scores. Z-scores are created by subtracting the mean and dividing by the standard deviation for each issue against the participants set. This ensures that variations in the range of participants' scoring is accounted for. We then ranked the average z-scores across the issues and selected the highest ranked 41 (approximately cutting the long list in half).

We discussed two potential reforms on the previous scoring approach: breaking scoring down across the three criteria, and including uncertainty estimates. We decided against both potential reforms. Experts are poor at estimating their own uncertainty and this could incentivise overconfidence. We decided that greater disaggregation in voting was likely to impose a greater burden on participants while providing little additional benefit. Moreover, keeping the protocol similar to the 2017 scan was desirable for comparison.

One amendment was made to the previous horizon scanning methodology: the introduction of "devil's advocates" into the process. Goodwin and Wright (2010) have noted that most forecasting methods are inadequate for identifying high-impact, low-probability events (some times called "black swan events"). However, the Delphi technique can be better suited to the task if it includes devil's advocates who can advocate for less likely but significant issues. We empowered two individuals during the first phase of the process to propose more speculative and transformative issues. Two different participants were then asked during the third phase (workshop deliberation) to provide more critical inputs and actively push against the prevailing, dominant view during discussions. In each case their designation was not revealed to the group.

The devil's advocates appear to have been a useful addition and were disproportionately successful in suggesting issues. Six of the nine issues they proposed in the first round made it through to the short list, and four of the six issues they proposed in the second round made it through to the final list of 20; with 38 participants, we would expect approximately only one issue for every second participant to make it through to the final list. 68% of participants had heard of the issues proposed by the devil's advocates, making these issues moderately more novel than the rest. Overall, an average of 70% of participants had heard of each issue. The level of novelty of the issues suggested by devil's advocates is partly skewed by two more well-known issues which both scored 82.35%. When both of these issues were excluded, the devil's advocates suggestions were significantly more novel at an average of 61%.

6.3 Phase three: Workshop preparation, deliberation and re-scoring

The 41 issues with the highest scores were kept as a part of a shortlist. These were sent back to participants on the 13 September 2019. Participants were assigned "cynic" roles for each issue. This involved doing deeper background research into the topic. Each issue had at least two cynics, ensuring that at least three participants (the cynics and proposer) had an in-depth knowledge of the area. The workshop was held in Cambridge on 9 October 2019 with 25 participants; 13 could not attend due to other obligations. This resulted in a group with approximately the same characteristics as the group that was involved in the first two phases. The characteristics of both groups are compared in Table 3. Overall, the gender balance was maintained (although the slight skew was reversed towards female participants), the disciplinary split between social and physical scientists was approximately the same, and the geographical coverage became less balanced due to the loss of participants from Peru, Zambia and Malaysia.

These discussions were overseen by an experienced facilitator (WJS, with LK and AR acting as scribes) and followed a deliberate structure. Each issue was discussed for approximately ten minutes before being voted on anonymously. During discussions, proposers of the issue were asked not to speak until at least three other respondents had contributed. This was done to avoid biasing the conversation and allowing the cynics time to provide an orientating, more neutral intervention. The standardised z-scores for each issue were calculated and ranked at the end of the workshop, resulting in a top 20 list. The decision to keep the list to 20 was made by consensus by the workshop group and was influenced by a significant difference between the z-scores of the top and bottom 20 issues, but a much smaller spread of scores within the top 20. Participants were then given time to discuss the final list and whether any amendments were needed. The group was content with the spread of the final list and that it accurately reflected the deliberations and hence decided that no alterations were needed.

Table 3: A comparative analysis of the groups involved with phases one and two, and phase three (the workshop).

Characteristics	Phases one and two	Phase three (workshop)
Sample Size	38	25
Gender Balance	21 male participants (55%) and 17 female participants (45%)	13 females (52%) and 12 males (48%).
Geographical Coverage	13 countries (UK, US, Canada, Australia, Germany, Croatia, Thailand, France, Chile, Peru, Switzerland, Malaysia, Zambia and Pakistan)	10 countries (UK, US, Canada, Australia, Germany, Croatia, Thailand, France, Chile, Switzerland, and Pakistan)
Disciplinary Distribution	15 participants from humanities and social sciences (39%), and 23 from natural sciences (61%).	9 participants from humanities and social sciences (36%) and 16 from natural sciences (64%).

Acknowledgements

We thank Andrew Balmer, George Church, Nick Matthews, Robert Meckin, Penny Polson, Fatemeh Salehi, and Andrew Watkins for their involvement in submitting issues.

This project was made possible through the support of a grant from Templeton World Charity Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of Templeton World Charity Foundation.

Philip Shapira acknowledges support from the Manchester Synthetic Biology Research Centre for Fine and Speciality Chemicals (grant BB/M017702/1), funded by the Biotechnology and Biological Sciences Research Council, and the Future Biomanufacturing Research Hub (grant EP/S01778X/1), funded by the Engineering and Physical Sciences Research Council and the Biotechnology and Biological Sciences Research Council as part of UK Research and Innovation.

Deborah Scott and Robert Smith acknowledge support from the European Research Council [ERC 616510].

Rainer Breitling, Philip Shapira and Eriko Takano acknowledge support from the Manchester Synthetic Biology Research Centre for Fine and Speciality Chemicals (grant BB/M017702/1), funded by the Biotechnology and Biological Sciences Research Council, and the Future Biomanufacturing Research Hub (grant EP/S01778X/1), funded by the Engineering and Physical Sciences Research Council and the Biotechnology and Biological Sciences Research Council as part of UK Research and Innovation.

Notes and References

[1] CSER's first horizon scan was performed in 2017 and was published as Wintle, B. C. et al. 'A Transatlantic perspective on 20 emerging issues in biological engineering', *eLife, 6* (2017): 1–21. https://doi.org/10.7554/eLife.30247

[2] Guston, D. H. 'Understanding "anticipatory governance"', *Social Studies of Science, 44* (2013): 218–42. https://doi.org/10.1177/0306312713508669

[3] Wintle et al. (2017).

[4] Hanea, A. M. et al. 'Investigate Discuss Estimate Aggregate for structured expert judgement', *International Journal of Forecasting, 33* (2017): 267–79. https://doi.org/10.1016/j.ijforecast.2016.02.008.

[5] Sutherland, W. J. and H. J. Woodroof 'The need for environmental horizon scanning', *Trends in Ecology and Evolution, 24*(10) (2009): 523–27. https://doi.org/10.1016/j.tree.2009.04.008

[6] Sutherland, W. J. et al. 'The identification of 100 ecological questions of high policy relevance in the UK', *Journal of Applied Ecology, 43*(4) (2006): 617–27. https://doi.org/10.1111/j.1365-2664.2006.01188.x; Sutherland, William J. et al. 'A 2017 horizon scan of emerging issues for global conservation and biological diversity', *Trends in Ecology and Evolution, 32*(1) (2017): 31–40. https://doi.org/10.1016/j.tree.2016.11.005

[7] Ricciardi, A. et al. 'Invasion science: A horizon scan of emerging challenges and opportunities', *Trends in Ecology and Evolution* (2017). https://doi.org/10.1016/j.tree.2017.03.007

8 Pretty, J. et al. 'The top 100 questions of importance to the future of global agriculture', *International Journal of Agricultural Sustainability, 8*(4) (2010): 219–36. https://doi.org/10.3763/ijas.2010.0534

9 Boddie, C. et al. 'Assessing the bioweapons threat', *Science, 349*(6250) (2015). https://doi.org/10.1126/science.aab0713

10 Sutherland et al. (2017).

11 Kennicutt, M. C. et al. 'Polar research: Six priorities for Antarctic science', *Nature, 512*(7512) (2014): 23–25. https://doi.org/10.1038/512023a

12 Kennicutt, M. et al. 'Sustained Antarctic research: A 21st century imperative', *One Earth* (2019). https://doi.org/10.1016/j.oneear.2019.08.014

13 Kennicutt, M. C. et al. 'A roadmap for Antarctic and Southern Ocean science for the next two decades and beyond', *Antarctic Science, 27*(1) (2015): 3–18. https://doi.org/10.1017/S0954102014000674

14 Cyranoski, D. 'The CRISPR-Baby scandal: What's next for human gene-editing', *Nature* (2019). https://doi.org/10.1038/d41586-019-00673-1

15 South, P. F. et al. 'Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field', *Science, 363*(6422) (2019): eaat9077. https://doi.org/10.1126/science.aat9077

16 WHO. *DRAFT Landscape of COVID-19 Candidate Vaccines — 26 April 2020*. World Health Organisation (2020).

17 Jessop-Fabre, M. M. and N. Sonnenschein. 'Improving reproducibility in synthetic biology', *Frontiers in Bioengineering and Biotechnology, 7*(18) (2019): 1–6. https://doi.org/10.3389/fbioe.2019.00018

18 Lentzos, F. and C. Invernizzi. *'Laboratories in the cloud', The Bulletin* (2019). https://thebulletin.org/2019/07/laboratories-in-the-cloud/

19 Palmer, M. J., F. Fukuyama and D. A. Relman. 'A more systematic approach to biological risk', *Science, 350*(6267) (2015): 1471–73. https://doi.org/10.1126/science.aad8849; Dunlap, G. and E. Pauwels. *The Intelligent and Connected Bio-Labs of the Future: Promise and Peril in the Fourth Industrial Revolution* (2019).

20 Altieri, M. A. et al. 'Agroecology and the design of climate change-resilient farming systems', *Agronomy for Sustainable Development, 35*(3) (2015): 869–90. https://doi.org/10.1007/s13593-015-0285-2

21 Dhankher, O. P. and C. H. Foyer. 'Climate resilient crops for improving global food security and safety', *Plant Cell and Environment, 41* (2018): 877–84. https://doi.org/10.1111/pce.13207

22 Nuccio, M. L. et al. 'Where are the drought tolerant crops? an assessment of more than two decades of plant biotechnology effort in crop improvement', *Plant Science, 273* (2018): 110–19. https://doi.org/10.1016/j.plantsci.2018.01.020

23 De La Concepcion, J. C. et al. 'Protein engineering expands the effector recognition profile of a rice NLR immune receptor', *eLife* (2019): 1–19. https://doi.org/10.7554/eLife.47713

24 Reynoso, M. A. et al. 'Evolutionary flexibility in flooding response circuitry in angiosperms', *Science, 365* (2019): 1291–95. https://doi.org/10.1126/science.aax8862

25 Feins, S. et al. 'An introduction to Chimeric Antigen Receptor (CAR) T-Cell immunotherapy for human cancer', *American Journal of Hematology, 94*(1) (2019): 3–9. https://doi.org/10.1002/ajh.25418

[26] AlQuraishi, M. 'End-to-end differentiable learning of protein structure', *Cell Systems, 8*(4) (2019): 292–301. https://doi.org/10.1016/j.cels.2019.03.006; Yang, K. K., Z. Wu and F. H. Arnold. 'Machine-learning-guided directed evolution for protein engineering', *Nature Methods, 16* (2019a): 687–94. https://doi.org/10.1038/s41592-019-0496-6

[27] Coutts. *United Kingdom 2017: Total Value of Million-Pound Donations Reaches Highest in 10 Years* (2019). https://philanthropy.coutts.com/en/reports/2017/united-kingdom/findings.html; Depecker, T., M.-O. Déplaude and N. Larchet. 'Philanthropy as an investment: Contribution to a study of reproduction and legitimation strategies of economic elites', *Politix, 121* (2018): 9–27.

[28] Reich, R. *Just Giving: Why Philanthropy Is Failing Democracy and How It Can Do Better*. Princeton University Press (2018).

[29] Rafols, I. and A. Yegros. *Is Research Responding to Health Needs*? (2018). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3106713

[30] Lentzos, F. 'Will splashy philanthropy cause the biosecurity field to focus on the wrong risks?', *Bulletin of the Atomic Scientists* (April 2019). https://thebulletin.org/2019/04/will-splashy-philanthropy-cause-the-biosecurity-field-to-focus-on-the-wrong-risks/

[31] Finnegan, T. and A. Hall. *Identification and Genomic Data* (2017). https://www.phgfoundation.org/documents/PHGF-Identification-and-genomic-data.pdf

[32] Wee, S.-L. 'American DNA expertise helps Beijing crack down', *New York Times* (2019). https://www.nytimes.com/2019/02/21/business/china-xinjiang-uighur-dna-thermo-fisher.html

[33] Haag, M. 'FamilyTreeDNA admits to sharing genetic data with FBI', *New York Times* (February 2019). https://www.nytimes.com/2019/02/04/business/family-tree-dna-fbi.html

[34] Solove, D. *Nothing to Hide: The False Tradeoff Between Privacy and Security* (1st ed.). Yale University Press (2011).

[35] Neve, P. 'Gene drive systems: Do they have a place in agricultural weed management?', *Pest Management Science, 74* (2018): 2671–79. https://doi.org/10.1002/ps.5137; Gantz, V. M. et al. 'Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi', PNAS, 112* (2015): E6736–E6743. https://doi.org/10.1073/pnas.1521077112

[36] Collins, J. P. 'Gene drives in our future: Challenges of and opportunities for using a self-sustaining technology in pest and vector management', *BMC Proceedings, 12*(9) (2018): 37–41. https://doi.org/10.1186/s12919-018-0110-4; Neve (2018).

[37] Evans, S. W. and M. J. Palmer. 'Anomaly handling and the politics of gene drives', *Journal of Responsible Innovation, 5*(1) (2018): S223–S242. https://doi.org/10.1080/23299460.2017.1407911

[38] Montenegro de Wit, M. 'Gene driving the farm: Who decides, who owns, and who benefits?', *Agroecology and Sustainable Food Systems, 43*(9) (2019): 1054–74. https://doi.org/10.1080/21683565.2019.1591566

[39] Wegrzyn, R. *Safe Genes, DARPA* (2019). https://www.darpa.mil/program/safe-genes

[40] Yang, X. et al. 'Bioinspired neuron-like electronics', *Nature Materials, 18* (2019b): 510–17. https://doi.org/10.1038/s41563-019-0292-9

[41] Yetisen, A. K. 'Biohacking', *Trends in Biotechnology, 36*(8) (2018): 744–47. https://doi.org/10.1016/j.tibtech.2018.02.011

42 Gallegos, J. E. et al. 'The Open Insulin Project: A case study for "biohacked" medicines', *Trends in Biotechnology, 36*(12) (2018): 1211–18. https://doi.org/10.1016/j.tibtech.2018.07.009

43 Munos, B. 'Can open-source drug r&d repower pharmaceutical innovation?', *Clinical Pharmacology & Therapeutics, 87*(5) (2010): 534–36; Gassmann, O. et al. 'The open innovation challenge: how to partner for innovation', *Leading Pharmaceutical Innovation*. Springer (2018), pp. 1–133. https://doi.org/10.1007/978-3-319-66833-8_6; Open Source Pharma. *Open Source Pharma: Medicine for All* (2020). https://www.opensourcepharma.net/

44 Blum, D. *The Poisoner's Handbook: Murder and the Birth of Forensic Medicine in Jazz Age New York*. Penguin Books (2010).

45 Coleman, K. and R. A. Zilinskas. 'Fake botox, real threat', *Scientific American, 306*(6) (2010): 84–89. https://doi.org/10.1038/scientificamerican0610-84

46 WHO. *Global Action Plan on Antimicrobial Resistance* (2015). https://www.who.int/antimicrobial-resistance/publications/global-action-plan/en/

47 Schmidt, C. 'Publisher correction: Phage therapy's latest makeover', *Nature Biotechnology, 37* (2019): 581–86. https://doi.org/10.1038/s41587-019-0158-3

48 Pires, D. P. et al. 'Genetically engineered phages: A review of advances over the last decade', *Microbiology and Molecular Biology Reviews, 80* (2016): 523–43. https://doi.org/10.1128/mmbr.00069–15

49 Japsen, B. 'Pfizer partners with IBM Watson to advance cancer drug discovery', *Forbes* (December 2016). https://www.forbes.com/sites/brucejapsen/2016/12/01/pfizer-partners-with-ibm-watson-to-advance-cancer-drug-discovery/#12d8a3a81b1e

50 iCarbonX. *The iCarbonX Difference* (2018). https://www.icarbonx.com/en/about.html

51 Stokes, J. M. et al. 'A deep learning approach to antibiotic discovery', *Cell, 180*(4) (2020): 688–702. https://doi.org/10.1016/j.cell.2020.01.021

52 Cho, H., D. J. Wu and B. Berger. 'Secure genome-wide association analysis using multiparty computation', *Nature Biotechnology, 36* (2018): 547–51. https://doi.org/10.1038/nbt.4108

53 Yakubu, A. et al. 'Model framework for governance of genomic research and biobanking in Africa — A content description', *AAS Open Research* (2018): 1–13. https://doi.org/10.12688/aasopenres.12844.2

54 Broaders, E., C. G. M. Gahan and J. R. Marchesi. 'Mobile genetic elements of the human gastrointestinal tract: Potential for spread of antibiotic resistance genes', *Gut Microbes, 4*(4) (2013): 271–80. https://doi.org/10.4161/gmic.24627

55 Foo, J. L. et al. 'Microbiome engineering: Current applications and its future', *Biotechnology Journal, 12*(3) (2017): 1–9. https://doi.org/10.1002/biot.201600099

56 Ronda, C. et al. 'Metagenomic engineering of the mammalian gut microbiome in situ', *Nature Methods, 16*(2) (2019): 167–70. https://doi.org/10.1038/s41592-018-0301-y

57 Fasani, E. et al. 'The potential of genetic engineering of plants for the remediation of soils contaminated with heavy metals', *Plant Cell and Environment, 41*(5) (2018): 1201–32. https://doi.org/10.1111/pce.12963

58 Basu, S. et al. 'Engineering pgpmos through gene editing and systems biology: A solution for phytoremediation?', *Trends in Biotechnology, 36*(5) (2018): 499–510. https://doi.org/10.1016/j.tibtech.2018.01.011

59 Merlin, M., M. Pezzotti and L. Avesani. 'Edible plants for oral delivery of biopharmaceuticals', *British Journal of Clinical Pharmacology, 83*(1) (2017): 71–81. https://doi.org/10.1111/bcp.12949

60 Tacket, C. O. et al. 'Human immune responses to a novel Norwalk virus vaccine delivered in transgenic potatoes', *The Journal of Infectious Diseases, 182*(1) (2000): 302–05. https://doi.org/10.1086/315653

61 Marsian, J. et al. 'Plant-made nervous necrosis virus-like particles protect fish against disease', *Frontiers in Plant Science, 10*(880) (2019): 1–11. https://doi.org/10.3389/fpls.2019.00880

62 Berardi, A. et al. 'Stability of plant virus-based nanocarriers in gastrointestinal fluids', *Nanoscale, 10*(4) (2018): 1667–79. https://doi.org/10.1039/c7nr07182e

63 Merlin et al. (2017).

64 PMC. *Personalised Medicine in Brief: Volume 12* (2019). http://www.personalizedmedicinecoalition.org/Userfiles/PMC-Corporate/file/PM_in_Brief_Vol_122.pdf

65 Bilkey, G. A. et al. 'Optimizing precision medicine for public health', *Frontiers in Public Health, 7*(42) (2019): 1–9. https://doi.org/10.3389/fpubh.2019.00042; Genetics Home Reference. *What are Some of the Challenges Facing Precision Medicine and the Precision Medicine Initiative*? (2019). https://ghr.nlm.nih.gov/primer/precisionmedicine/challenges

66 European Commission. *Commission Expert Group on Bio-based Products: Final Report* (2017). http://bio-based.eu/downloads/commission-expert-group-on-bio-based-products-final-report/

67 Matthews, N. E., L. Stamford and P. Shapira. 'Aligning sustainability assessment with responsible research and innovation: Towards a framework for constructive sustainability assessment', *Sustainable Production and Consumption, 20* (2019): 58–73. https://doi.org/10.1016/j.spc.2019.05.002

68 HM Government. *Industrial Strategy: Growing the Bioeconomy* (2018). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/761856/181205_BEIS_Growing_the_Bioeconomy__Web_SP_.pdf

69 Bruce, T. J. A. et al. 'The first crop plant genetically engineered to release an insect pheromone for defence', *Scientific Reports, 5*(11183) (2015): 19. https://doi.org/10.1038/srep11183

70 Ding, B. J. et al. 'A plant factory for moth pheromone production', *Nature Communications, 25*(5) (2014): 3353. https://doi.org/10.1038/ncomms4353

71 Ward, K. D. *Statement by Kenneth D Ward* (2019). https://www.opcw.org/sites/default/files/documents/2019/07/ec91nat08%28e%29.pdf

72 NIH. *Why Is the BRAIN Initiative Needed? BRAIN Initiative* (2019). https://braininitiative.nih.gov/about/overview

73 Nixdorff, K. et al. 'Dual-use nano-neurotechnology', *Politics and the Life Sciences, 37*(2) (2018): 180–202. https://doi.org/10.1017/pls.2018.15

74 Erb, T. J. 'Back to the future: Why we need enzymology to build a synthetic metabolism of the future', *Beilstein Journal of Organic Chemistry, 15* (2019): 551–57. https://doi.org/10.3762/bjoc.15.49

75 Trudeau, D. L. et al. 'Design and in vitro realization of carbon-conserving photorespiration', *PNAS, 115* (2018): E11455–E11464. https://doi.org/10.1073/pnas.1812605115

76 Schwander, T. et al. (2016) 'A synthetic pathway for the fixation of carbon dioxide in vitro', *Science, 354*(6314) (2016): 900–04. https://doi.org/10.1126/science.aah5237

77 Gleizer, S. et al. 'Conversion of Escherichia coli to generate all biomass carbon from CO2', *Cell, 179*(6) (2019): 1255–63.

78 Niu, D. et al. 'Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9', *Science, 357*(6357) (2017): 1303–7. https://doi.org/10.1126/science.aan4187

79 UNOS. *Transplant Trends, United Network for Organ Sharing* (2019). https://unos.org/data/transplant-trends/

80 Nuffield Council on Bioethics. *Animal-to-HumanTransplants:TheEthicsofXenotransplantation* (1996).

81 Fond, G. et al. 'Innovative mechanisms of action for pharmaceutical cognitive enhancement: A systematic review', *Psychiatry Research, 229*(1–2) (2015): 12–20. https://doi.org/10.1016/j.psychres.2015.07.006

82 Wexler, A. 'The social context of "do-it-yourself" brain stimulation: Neurohackers, biohackers, and lifehackers', *Frontiers in Human Neuroscience, 11*(224) (2017): 1–6. https://doi.org/10.3389/fnhum.2017.00224

83 Maslen, H. et al. 'The regulation of cognitive enhancement devices: Refining Maslen et al.'s model', *Journal of Law and the Biosciences, 2*(3) (2016): 754–67. https://doi.org/10.1093/jlb/lsv029

84 See, for example, Licurse, A. et al. 'The impact of disclosing financial ties in research and clinical care: A systematic review', *Archives of Internal Medicine, 170*(8) (2010): 675–82. https://doi.org/10.1001/archinternmed.2010.39

85 Evans, B. R. et al. 'Transgenic aedes aegypti mosquitoes transfer genes into a natural population', *Scientific Reports, 9*(1) (2019): 1–6. https://doi.org/10.1038/s41598-019-49660-6

86 Peters, G. P. et al. 'Global carbon budget 2019', *Earth System Science Data, 11* (October 2019): 11783–838.

87 Steffen, W. et al. 'Trajectories of the Earth system in the Anthropocene', *PNAS, 115* (2018): 8252–59. https://doi.org/10.1073/pnas.1810141115; Lenton, T. et al. 'Climate tipping points — Too risky to bet against', *Nature, 575* (2019): 592–95.

88 Sackman, H. *Delphi Critique: Expert Opinion, Forecasting and Group Process*. Lexington Books (1975).

89 Hanea et al. (2017).

90 de Loë, R. C. et al. 'Advancing the state of Policy Delphi Practice: A systematic review evaluating methodological evolution, innovation, and opportunities', *Technological Forecasting and Social Change, 104* (2016): 78–88. https://doi.org/10.1016/j.techfore.2015.12.009

91 Parente, R. and J. Anderson-Parente. 'A case study of long-term Delphi Accuracy', *Technological Forecasting and Social Change, 78*(9) (2011): 1705–11. https://doi.org/10.1016/j.techfore.2011.07.005

92 Hanea, A. M. et al. ‘Classical meets modern in the IDEA protocol for structured expert judgement’, *Journal of Risk Research, 21*(4) (2018): 417–33. https://doi.org/10.1080/13669877.2016.1215346; Hanea, Anca M. et al. ‘The value of performance weights and discussion in aggregated expert judgments’, *Risk Analysis, 38*(9) (2018): 1781–94. https://doi.org/10.1111/risa.12992

93 Hanea et al. (2017).

94 Hemming, V. et al. ‘Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management, *PLoS ONE* (2018). https://doi.org/10.1371/journal.pone.0198468

95 Barons, M. J. et al. ‘Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation’, *Journal of Apicultural Research, 57*(5) (2018). 593–604. https://doi.org/10.1080/00218839.2018.1494891

17. Artificial Canaries: Early Warning Signs for Anticipatory and Democratic Governance of AI

Carla Zoe Cremer and Jess Whittlestone

Highlights:

- This chapter proposes a method for identifying early warning signs of transformative progress in Artificial Intelligence (AI), and discusses how these can support the anticipatory and democratic governance of AI. These early warning signs are called "canaries", based on the use of canaries to provide early warnings of unsafe air pollution in coal mines.
- The author's method combines expert elicitation and collaborative causal graphs to identify key milestones and the relationships between them. They present two illustrations of how this method could be used: to identify early warnings of harmful impacts of language models on political systems; and of progress towards high-level machine intelligence.
- Identifying early warning signs of transformative applications can support more efficient monitoring and timely regulation of progress in AI: as AI advances, its impacts on society may be too great to be governed retrospectively.
- It is essential that those impacted by AI have a say in how it is governed. Early warnings can give the public time and focus to influence emerging technologies using democratic, participatory processes.

https://doi.org/10.11647/OBP.0360.17

This chapter was originally published in 2021 the *International Journal of Interactive Multimedia & Artificial Intelligence*. Like other contributions to this volume, it proposes the use of expert elicitation and the collaborative development of knowledge and understanding, in this case through the use of causal graphs. Methodological comparisons can be explored by reviewing Chapters 7, 8 or 16, whilst the core arguments concerning representation and democracy are also examined in different ways in Chapter 2 and 22.

I. Introduction

Progress in Artificial Intelligence (AI) research has accelerated in recent years. Applications are already changing society[1] and some researchers warn that continued progress could precipitate transformative impacts.[2] We use the term "transformative AI" to describe a range of possible advances with potential to impact society in significant and hard-to-reverse ways.[3] For example, future machine learning systems could be used to optimise management of safety-critical infrastructure.[4] Advanced language models could be used in ways that corrupt our online information ecosystem[5] and future advances in AI systems could trigger widespread labour automation.[6]

There is an urgent need to develop anticipatory governance approaches to AI development and deployment. As AI advances, its impacts on society will become more profound, and some harms may be too great to rely on purely "reactive" or retrospective governance.

Anticipating future impacts is a challenging task. Experts show substantial disagreement about when different advances in AI capabilities should be expected.[7] Policy-makers face challenges in keeping pace with technological progress: it is difficult to foresee impacts before a technology is deployed, but after deployment it may already be too late to shape impacts, and some harm may already have been done.[8] Ideally, we would focus preventative, anticipatory efforts on applications which are close enough to deployment to be meaningfully influenced today, but whose impacts we are not already seeing. Finding "early warning signs" of transformative AI applications can help us to do this.

Early warning signs can also help democratise AI development and governance. They can provide time and direction for much-needed public discourse about what we want and do not want from AI. It is not enough for anticipatory governance to look out for supposedly "inevitable" future impacts. We are not mere bystanders in this AI revolution: the futures we occupy will be futures of our own making, driven by the actions of technology developers, policy-makers, civil society and the public. In order to prevent foreseeable harms towards those people who bear the effects of AI deployments, we must find ways for AI developers to be held accountable to the society which they are embedded in. If we want AI to benefit society broadly, we must urgently find ways to give democratic control to those who will be impacted. Our aim with identifying early warning signs is to develop anticipatory methods which can prompt a focussed civic discourse around significant developments and provide a wider range of people with the information they need to contribute to conversations about the future of AI.

We present a methodology for identifying early warning signs of potentially transformative impacts of AI and discuss how these can feed into more anticipatory and democratic governance processes. We call these early warning signs "canaries" based on the practice of using canaries to provide early warnings of unsafe air pollution in coal mines in the industrial revolution. Others before us have used this term in the context of AI to stress the importance of early warning signs[9] but this is the first attempt to outline in detail how such "artificial canaries" might be identified and used.

Our methodology is a prototype but we believe it provides an important first step towards assessing and then trialling the feasibility of identifying canaries. We first present the approach and then illustrate it on two high-level examples, in which we identify preliminary warning signs of AI applications that could undermine democracy, and warning signs of progress towards High-Level Machine Intelligence (HLMI). We explain why early warning signs are needed by drawing on the literature of Participatory Technology Assessments, and we discuss the advantages and practical challenges of this method in the hope of preparing future research that might attempt to put this method into practise. Our theoretical exploration of a method to identify early warning signs of transformative applications provides a foundation towards more anticipatory, accountable and democratic governance of AI in practice.

2. Related Work

We rely on two main bodies of work. Our methodology for identifying canaries relies on the literature on *forecasting and monitoring AI*. Our suggestions for how canaries might be used once identified build on work on *Participatory Technology Assessments*, which stresses a more inclusive approach to technology governance. While substantial research exists in both these areas, we believe this is the first piece of work that shows how they could feed into each other.

A. AI forecasting and monitoring

Over the past decade, an increasing number of studies have attempted to forecast AI progress. They commonly use expert elicitations to generate probabilistic estimates for when different AI advances and milestones will be achieved.[10] For example, Baum et al. ask experts about when specific milestones in AI will be achieved, including passing the Turing Test or passing third grade.[11] Both Müller and Bostrom[12] and Grace et al.[13] ask experts to predict the arrival of high-level machine intelligence (HLMI), which the latter define as when "unaided machines can accomplish every task better and more cheaply than human workers".

However, we should be cautious about giving results from these surveys too much weight. These studies have several limitations, including the fact that the questions asked are often ambiguous, that expertise is narrowly defined, and that respondents do not receive training in quantitative forecasting.[14] Experts disagree substantially about when crucial capabilities will be achieved,[15] but these surveys cannot tell us who (if anyone) is more accurate in their predictions.

Issues of accuracy and reliability aside, forecasts focused solely on timelines for specific events are limited in how much they can inform our decisions about AI today. While it is interesting to know how much experts disagree on AI progress via these probabilistic estimates, they cannot tell us why experts disagree or what would change their minds. Surveys tell us little about what early warning signs to look out for or where we should place our focus today to shape the future development and impact of AI.

At the same time, several projects have begun to track and measure progress in AI.[16] These projects focus on a range of indicators relevant to

AI progress, but do not make any systematic attempt to identify which markers of progress are more important than others for the preparation of transformative applications. Time and attention for tracking progress is limited and it would be helpful if we were able to prioritise and monitor those research areas that are most relevant to mitigating risks.

Recognising some of the limitations of existing work, Gruetzemacher aims for a more holistic approach to AI forecasting.[17] This framework emphasises the use of the Delphi technique[18] to aggregate different perspectives of a group of experts, and cognitive mapping methods to study how different milestones relate to one another, rather than to simply forecast milestones in isolation. We agree that such methods might address some limitations of previous work in both AI forecasting and monitoring. AI forecasting has focused on timelines for particularly extreme events, but these timelines are subject to enormous uncertainty and do not indicate near-term warning signs. AI measurement initiatives have the opposite limitation: they focus on near-term progress, but with little systematic reflection on which avenues of progress are, from a governance perspective, more important to monitor than others. What is needed are attempts to identify areas of progress today that may be particularly important to pay attention to, given concerns about the kinds of transformative AI systems that may be possible in future.

B. Participatory Technology Assessments

Presently, the impacts of AI are largely shaped by a small group of powerful people with a narrow perspective which can be at odds with public interest.[19] Only a few powerful actors, such as governments, defence agencies, and firms the size of Google or Amazon, have the resources to conduct ambitious research projects. Democratic control over these research projects is limited. Governments retain discretion over what gets regulated, large technology firms can distort and avoid policies via intensive lobbying[20] and defence agencies may classify ongoing research.

Recognising these problems, a number of initiatives over the past few years have emphasised the need for wider participation in the development and governance of AI.[21] In considering how best to achieve this, it is helpful to look to the field of science and technology studies

(STS) which has long considered the value of democratising research progress.[22] Several publications refer to the "participatory turn" in STS[23] and an increasing interest in the role of the non-expert in technology development and assessment.[24] More recently, in the spirit of "democratic experimentation",[25] various methods for civic participation have been developed and trialled, including deliberative polls, citizen juries and scenario exercises.[26]

With a widening conception of expertise, a large body of research on "participatory technology assessment" (PTA) has emerged, aiming to examine how we might increase civic participation in how technology is developed, assessed and rolled out. We cannot summarise this wide-ranging and complex body of work fully here. But we point towards some relevant pieces for interested readers to begin with. Biegelbauer and Loeber[27] and Rowe and Frewer[28] present a typology of the methods and goals of participating, which now come in many forms. This means that assessments of the success of PTAs are challenging[29] and ongoing because different studies evaluate different PTA processes against different goals.[30] Yet while scholars recognise remaining limitations of PTAs,[31] several arguments for their advantages have been brought forward, ranging from citizen agency to consensus identification and justice. There are good reasons to believe that non-experts possess relevant end-user expertise. They often quickly develop the relevant subject-matter understanding to contribute meaningfully, leading to better epistemic outcomes due to a greater diversity of views which result in a cancellation of errors.[32] To assess the performance of PTAs, scholars draw from case studies and identify best practices.[33]

There is an important difference between truly participatory, democratically minded, technology assessments, and consultations that use the public to help legitimise a preconceived technology.[34] The question of how to make PTAs count in established representational democracies is an ongoing challenge to the field.[35] But Hsaio et al., who present a recent example of collective technology policy-making, show that success and impact with PTAs is possible.[36] Rask et al. draw from 38 international case studies to extract best practices,[37] building on Joss and Bellucci,[38] who showcase great diversity of possible ways in which to draw on the public. Comparing different approaches is difficult, but has been done.[39] Burgess and Chilvers present a conceptual framework with which to design and assess PTAs,[40] Ertiö et al. compare online versus offline methodologies[41] and in Rowe and Frewer we find a typology of various design choices for public engagement mechanisms.[42] See also, Rask for a helpful discussion on how to determine the diversity of participants;[43] Mauksch et al. on what counts as expertise in foresight;[44] and Lengwiler,[45] Chilvers,[46] and Saldivar et al.[47] for challenges to be aware of in implementing PTAs.

Many before us have noted that we need wider participation in the development and governance of AI, including by calling for the use of PTAs in designing algorithms.[48] We see a need to go beyond greater participation in addressing existing problems with algorithms and propose that wider participation should also be considered in conversations about future AI impacts.

Experts and citizens each have a role to play in ensuring that AI governance is informed by and inclusive of a wide range of knowledge, concerns and perspectives. However, the question of how best to marry expert foresight and citizen engagement is a challenging one. While a full answer to this question is beyond the scope of this chapter, what we do offer is a first step: a proposal for how expert elicitation can be used to identify important warnings which can later be used to facilitate timely democratic debate. For such debates to be useful, we first need an idea of which developments on the horizon can be meaningfully assessed and influenced, for which it makes sense to draw on public expertise and limited attention. This is precisely what our method aims to provide.

3. Identifying Early Warning Signs

We believe that identifying canaries for transformative AI is a tractable problem and worth investing research effort in today. Engineering and cognitive development present a proof of principle: capabilities are achieved sequentially, meaning that there are often key underlying capabilities which, if attained, unlock progress in many other areas. For example, musical protolanguage is thought to have enabled grammatical competence in the development of language in *homo sapiens*.[49] AI progress so far has also seen such amplifiers: the use of multi-layered non-linear learning or stochastic gradient descent arguably laid the foundation for unexpectedly fast progress on image recognition, translation and speech recognition.[50] By mapping out the dependencies between different capabilities needed to reach some notion of transformative AI, therefore, we should be able to identify milestones which are particularly important for enabling many others — these are our canaries.

The proposed methodology is intended to be highly adaptable and can be used to identify canaries for a number of important potentially transformative events, such as foundational research breakthroughs or the automation of tasks that affect a wide range of jobs. Many types of indicators could be of interest and classed as canaries, including algorithmic innovation that supports key cognitive faculties (e.g. natural language understanding); overcoming known technical challenges (such as improving the data efficiency of deep learning algorithms); or improved applicability of AI to economically-relevant tasks (e.g. text summarisation).

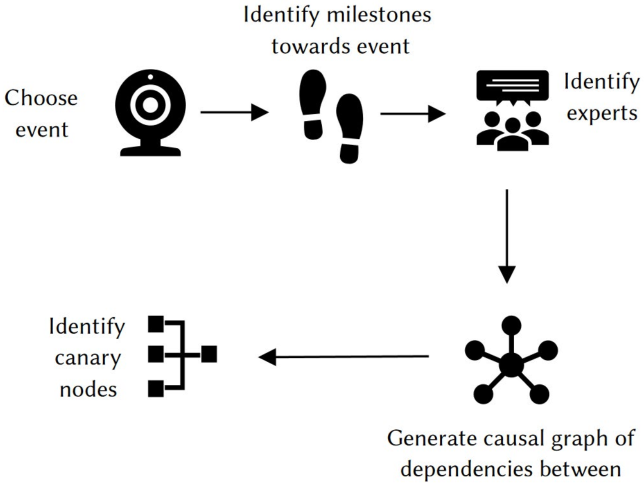

Identify milestones towards event
Choose event
Identify experts
Identify canary nodes
Generate causal graph of dependencies between

Fig. 1: Illustration of methodological steps to identify canaries of AI progress.

Given an event for which we wish to identify canaries, our methodology has three essential steps: (1) identifying key milestones towards the event; (2) identifying dependency relations between these milestones; and (3) identifying milestones which underpin many others as canaries. See Fig. 1 for an illustration. We here deliberately refrain from describing the method with too much specificity, because we want to stress the flexibility of our approach, and recognise that there is currently no one-size-fits-all approach to forecasting. The method will require adaptation to the particular transformative event in question, but each step of this method is suited for such specifications. We outline example adaptations of the method to particular cases.

A. Identifying milestones via expert elicitation

The first step of our methodology involves using traditional approaches in expert elicitation to identify milestones that may be relevant to the transformative event in question. Which experts are selected is crucial to the outcome and reliability of studies in AI forecasting. There are unavoidable limitations of using any form of subjective judgement in forecasting, but these limitations can be minimised by carefully thinking through the group selection. Both the direct expertise of individuals, and how they contribute to the diversity of the overall group, must be considered. See Mauksch et al. for a discussion of who counts as an expert in forecasting.[51]

Researchers should decide in advance what kinds of expertise are most relevant and must be combined to study the milestones that relate to the transformative event. Milestones might include technical limitations of current methods (e.g. adversarial attacks) and informed speculation about future capabilities (e.g. common sense) that may be important prerequisites to the transformative event. Consulting across a wide range of academic disciplines to order such diverse milestones is important. For example, a cohort of experts identifying and ordering milestones towards HLMI should include not only experts in machine learning and computer science but also cognitive scientists, philosophers, developmental psychologists, evolutionary biologists, or animal cognition experts. Such a group combines expertise on current capabilities in AI, with expertise on key pillars of cognitive development and the order in which cognitive faculties develop in animals. Groups which are diverse (on multiple dimensions) are expected to produce better epistemic outcomes.[52]

We encourage the careful design and phrasing of questions to enable participants to make use of their expertise, but refrain from demanding answers that lie outside their area of expertise. For example, asking machine learning researchers directly for milestones towards HLMI does not draw on their expertise. But asking machine learning researchers about the limitations of the methods they use every day — or asking psychologists what human capacities they see lacking in machines today — draws directly on their day-to-day experience.

Perceived limitations can then be transformed into milestones.

There are several different methods available for expert elicitation including surveys, interviews, workshops and focus groups, each with advantages and disadvantages. Interviews provide greater opportunity to tailor questions to the specific expert, but can be time-intensive compared to surveys and reduce the sample size of experts. If possible, some combination of the two may be ideal: using carefully selected semi-structured interviews to elicit initial milestones, followed-up with surveys with a much broader group to validate which milestones are widely accepted as being key.

B. Mapping causal relations between milestones

The second step of our methodology involves convening experts to identify causal relations between identified milestones: that is, how milestones may underpin, depend on, or affect progress towards other milestones. Experts should be guided in generating directed causal graphs, a type of cognitive map that elicits a person's perceived causal relations between components. Causal graphs use arrows to represent perceived causal relations between nodes, which in this case are milestones.[53]

This process primarily focuses on finding out whether or not a relationship exists at all; how precisely this relationship is specified can be adapted to the goals of the study. An arrow from A to B at minimum indicates that progress on A will allow for further progress on B. But this relationship can also be made more precise: in some cases indicating that progress on AI is *necessary* for progress on B, for example. The relationship between nodes may be either linear or nonlinear; again, this can be specified more precisely if needed or known.

Constructing and debating causal graphs can "help groups to convert tacit knowledge into explicit knowledge".[54] Causal graphs are used as decision support for individuals or groups, and are often used to solve problems in policy and management involving complex relationships between components in a system by tapping into experts' mental models and intuitions. We therefore suggest that causal graphs are particularly well-suited to eliciting experts' models and assumptions about the relationship between different milestones in AI development.

As a method, causal graphs are highly flexible and can be adapted to the preferred level of detail for a given study: they can be varied in complexity and can be analysed both quantitatively and qualitatively.[55] We neither exclude nor favour quantitative approaches here, due to the complexity and uncertainty of the questions around transformative events. Particularly for very high-level questions, quantitative approaches might not offer much advantage and might communicate a false sense of certainty. In narrower domains where there is more existing evidence, however, quantitative approaches may help to represent differences in the strength of relationships between milestones.

Eden notes that there are no ready-made designs that will fit all studies: design and analysis of causal mapping procedures must be matched to a clear theoretical context and the goal of the study.[56] We highlight a number of different design choices which can be used to adapt the process. As more studies use causal graphs in expert elicitations about AI developments, we can learn from the success of different design choices over time and identify best practices.

Scavarda et al. stress that interviews or collective brainstorming are the most accepted method for generating the data upon which to analyse causal relations.[57] Ackerman, Bryson, and Eden list heuristics on how to manage the procedure of combining graphs by different participants,[58] or see Montibeller and Belton for a discussion on evaluating different options presented by experts.[59] Scavarda et al. suggests visual, interactive tools to aid the process.[60] Eden[61] and Eden et al. [62] discuss approaches to analysing graphs and extracting the emergent properties, significant "core" nodes as well as hierarchical clusters. Core or "potent" nodes are those that relate to many clusters in the graphs and thus have implications for connected nodes. In our proposed methodology, such potent nodes play a central role in pointing to canary milestones. For more detail on the many options on how to generate, analyse and use causal graphs we refer the reader to the volume of Ackerman, Bryson, and Eden,[63] or reviews such as Scavardia et al. (2004 and 2006).[64] See Eden and Ackerman for an example of applying cognitive mapping to expert views on UK public policies,[65] and Ackerman and Eden for group problem-solving with causal graphs.[66]

We propose that identified experts be given instruction in generating either an individual causal graph, after which a mediated discussion between experts generates a shared graph; or that the groups of experts as a whole generate the causal graph via argumentation, visualisations and voting procedures if necessary. As Eden emphasises, any group of experts will have both shared and conflicting assumptions, which causal graphs aim to integrate in a way that approaches greater accuracy than that contained in any single expert viewpoint.[67] The researchers are free to add as much detail to the final maps as required or desired. Each node can be broken into subcomponents or justified with extensive literature reviews.

C. Identifying canaries

Finally, the resulting causal graphs can be used to identify nodes of particular relevance for progress towards the transformative event in question. This can be a node with a high number of outgoing arrows, i.e. milestones which unlock many others that are prerequisites for the event in question. It can also be a node which functions as a bottleneck — a single dependency node that restricts access to a subsequent highly significant milestone. See Fig. 2 for an illustration. Progress on these milestones can thus represent a "canary", indicating that further advances in subsequent milestones will become possible and more likely. These canaries can act as early warning signs for potentially rapid and discontinuous progress, or may signal that applications are becoming ready for deployment. Experts identify nodes which unlock or provide a bottleneck for a significant number of other nodes (some amount of discretion from the experts/conveners will be needed to determine what counts as "significant").

Of course, in some cases generating these causal graphs and using them to identify canaries may be as complicated as a full scientific research project. The difficulty of estimating causal relationships between future technological advances must not be underestimated. However, we believe it to be the case that each individual researcher already does this to some extent, when they chose to prioritise a research project, idea or method over another within a research paradigm. Scientists also debate the most fruitful and promising research avenues and arguably place bets on implicit maps of milestones as they pick a research agenda. The idea is not to generate maps that provide a perfectly accurate indication of warning signs, but to use the wisdom of crowds to make implicit assumptions explicit, creating the best possible estimate of which milestones may provide important indications of future transformative progress.

4. Using Early Warning Signs

Once identified, canary milestones can immediately help to focus existing efforts in forecasting and anticipatory governance. Given limited resources, early warning signs can direct governance attention to areas of AI progress which are soon likely to impact society and which can be influenced now. For example, if progress in a specific area of NLP (e.g. sentiment analysis) serves as a warning sign for the deployment of more engaging social bots to manipulate voters, policy-makers and regulators can monitor or regulate access and research on this research area within NLP.

We can also establish research and policy initiatives to monitor and forecast progress towards canaries. Initiatives might automate the collection, tracking and flagging of new publications relevant to canary capabilities, and build a database of relevant publications. They might use prediction platforms to enable collective forecasting of progress towards canary capabilities. Foundational research can try to validate hypothesised relationships between milestones or illuminate the societal implications of different milestones.

These forecasting and tracking initiatives can be used to improve policy prioritisation more broadly. For example, if we begin to see substantial progress in an area of AI likely to impact jobs in a particular domain, policy-makers can begin preparing for potential unemployment in that sector with greater urgency.

However, we believe the value of early warning signs can go further and support us in democratising the development and deployment of AI. Providing opportunities for participation and control over policy is a fundamental part of living in a democratic society. It may be especially important in the case of AI, since its deployment might indeed transform society across many sectors. If AI applications are to bring benefits across such wide-ranging contexts, AI deployment strategies must consider and be directed by the diverse interests found across those sectors. Interests which are underrepresented at technology firms are otherwise likely to bear the negative impacts.

There is currently an information asymmetry between those developing AI and those impacted by it. Citizens need better information about specific developments and impacts which might affect them. Public attention and funding for deliberation processes is not unlimited, so we need to think carefully about which technologies to direct public attention and funding towards. Identifying early warning signs can help address this issue, by focusing the attention of public debate and directing funding towards deliberation practises that centre around technological advancements on the horizon.

We believe early warning signs may be particularly well-suited to feed into Participatory Technology Assessments (PTAs), as introduced earlier. Early warning signs can provide a concrete focal point for citizens and domain experts to collectively discuss concerns. Having identified a specific warning sign, various PTA formats could be suited to consult citizens who are especially likely to be impacted. PTAs come in many forms and a full analysis of which design is best suited to assessing particular AI applications is beyond the scope of this article. But the options are plenty and PTAs show much potential (see Section 2). For example, Taiwan has had remarkable success and engagement with an open consultation of citizens on complex technology policy questions.[68] An impact assessment of PTA is not a simple task, but we hypothesise that carefully designed, inclusive PTAs would present a great improvement over how AI is currently developed, deployed and governed. Our suggestion is not limited to governmental bodies. PTAs or other deliberative processes can be run by research groups and private institutions such as AI labs, technology companies and think tanks who are concerned with ensuring AI benefits all of humanity.

5. Method Illustrations

We outline two examples of how this methodology could be adapted and implemented: one focused on identifying warning signs of a particular societal impact, the other on warning signs of progress towards particular technical capabilities. Both these examples pertain to high-level, complex questions about the future development and impacts of AI, meaning our discussion can only begin to illustrate what the process of identifying canaries would look like, and what questions such a process might raise. Since the results are only the suggestions of the authors of this chapter, we do not show a full implementation of the method whose value lies in letting a group of experts deliberate. As mentioned previously, the work of generating these causal maps will often be a research project of its own, and we will return later to the question of what level of detail and certainty is needed to make the resulting graphs useful.

A. First illustration: AI applications in voter manipulation

We show how our method could identify warning signs of the kind of algorithmic progress which could improve the effectiveness of, or reduce the cost of, algorithmic election manipulation. The use of algorithms in attempts to manipulate election results incur great risk for the epistemic resilience of democratic countries.[69]

Manipulations of public opinion by national and commercial actors are not a new phenomenon. We detail the history of how newly emerging technologies are often used for this purpose. [70] But recent advances in deep learning techniques, as well as the widespread use of social media, have introduced easy and more effective mechanisms for influencing opinions and behaviour. Several studies detail the various ways in which political and commercial actors incur harm to the information ecosystem via the use of algorithms.[71] Manipulators profile voters to identify susceptible targets on social media, distribute micro-targeted advertising, spread misinformation about policies of the opposing candidate and try to convince unwanted voters not to vote. Automation plays a large role in influencing online public discourse. Like et al.[72] and Ferrara[73] also note that manipulators use both human-run accounts and bots[74] or a combination of the two.[75] Misinformation[76] and targeted messaging[77] can have transformative implications for the resilience of democracies and the very possibility of collective action.[78]

Despite attempts by national and sub-national actors to apply algorithms to influence elections, their impact so far has been contested.[79] Yet foreign actors and national political campaigns will continue to have incentives and substantial resources to invest in such campaigns, suggesting their efforts are unlikely to wane in future. We may thus inquire what kinds of technological progress would increase the risk that elections can be successfully manipulated. We can begin this inquiry by identifying what technological barriers currently prevent full-scale election manipulation.

We would identify those technological limitations by drawing on the expertise of actors who are directly affected by these bottlenecks. Those might be managers of online political campaigns and foreign consulting firms (as described in Howard),[80] who specialise in influencing public opinion via social media, or governmental organisations across the world who comment on posts, target individual influencers and operate fake accounts to uphold and spread particular beliefs. People who run such political cyber campaigns have knowledge of what technological bottlenecks still constrain their influence on voter decisions. We recommend running a series of interviews to collect a list of limitations.

This list might include, for example, that the natural language functionality of social bots is a major bottleneck for effective online influence (for the plausibility of this being an important technical factor, see Howard).[81] Targeted users often disengage from a chat conversation after detecting that they are exchanging messages with social bots. Low retention time is presumably a bottleneck for further manipulation, which suggests that improvements in Natural Language Processing (NLP) would significantly reduce the cost of manipulation as social bots become more effective.

We will assume, for the purpose of this illustration that NLP were to be identified as a key bottleneck. We would then seek to gather experts (e.g. in a workshop) who can identify and map milestones (or current limitations) in NLP likely to be relevant to improving the functionality of social bots. This will include machine learning experts who specialise in NLP and understand the technical barriers to developing more convincing social bots, as well as experts in developmental linguistics and evolutionary biology, who can determine suitable benchmarks and the required skills, and who understand the order in which linguistic skills are usually developed in animals.

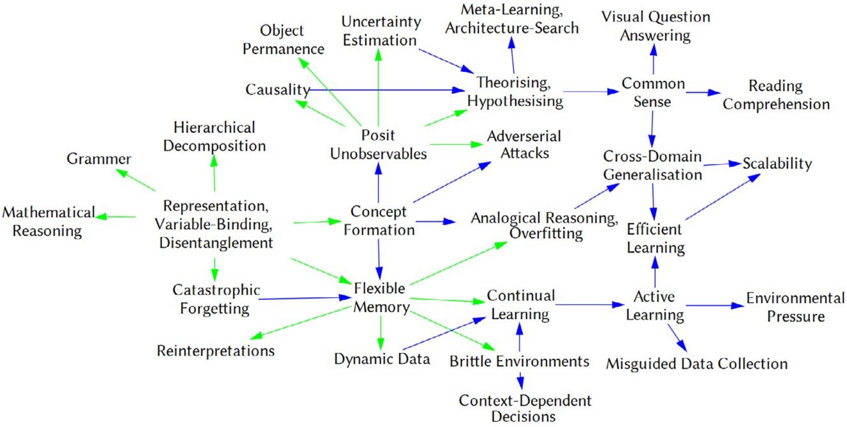


Fig. 2: Cognitive map of dependencies between milestones collected in expert elicitations. Arrows coloured in green signify those milestones that have most outgoing arrows. See appendix for description of each milestone and dependency relations between one "canary" node and subsequent nodes.

From these expert elicitation processes we would acquire a list of milestones in NLP which, if achieved, would likely lower the cost and increase the effectiveness of online manipulation. Experts would then order milestones into a causal graph of dependencies. Given the interdisciplinary nature of the question at hand, we suggest in this case that the graph should be directly developed by the whole group. A mediated discussion in a workshop context can help to draw out different connections between milestones and the reasoning behind them, ensuring participants do not make judgements outside their range of expertise. A voting procedure such as majority voting should be used if no consensus can be reached. In a final step, experts can highlight milestone nodes in the final graph which are either marked by many outgoing nodes or are bottlenecks for a series of subsequent nodes that are not accessed by an alternative pathway. These (e.g. sentiment analysis) are our canaries: areas of progress which serve as a warning sign of NLP being applied more effectively in voter manipulation.

Having looked at how this methodology can be used to identify warning signs of a specific societal impact, we next illustrate a different application of the method in which we aim to identify warning signs of a research breakthrough.

B. Second illustration: High-level Machine intelligence

We use this second example to illustrate in more detail what the process of developing a causal map might look like once initial milestones have been identified, and how canary capabilities can be identified from the map.

We define High-Level Machine Intelligence (HLMI) as an AI system (or collection of AI systems) that performs at the level of an average human adult on key cognitive measures required for economically relevant tasks. We choose to focus on HLMI since it is a milestone which has been the focus of previous forecasting studies[82] and which, despite the ambiguity and uncertain nature of the concepts, is interesting to attempt to examine, because it is likely to precipitate widely transformative societal impacts.

To trial this method, we used interview results from Cremer (2021).[83] 25 experts from a diverse set of disciplines (including computer science, cognitive science and neuroscience) were interviewed and asked what they believed to be the main limitations preventing current machine learning methods from achieving the capabilities of HLMI. These limitations can be translated into “milestones”: capabilities experts believe machine learning methods need to achieve on the path to HLMI, i.e. the output of Step 1 of our methodology.

Having identified key milestones, Step 2 of our methodology involves exploring dependencies between them using causal graphs. We use the software VenSim to illustrate hypothesised relationships between milestones (see Fig. 2). For example, we hypothesise that the ability to formulate, comprehend and manipulate abstract concepts may be an important prerequisite to the ability to account for unobservable phenomena, which is in turn important for reasoning about causality. This map of causal relations and dependencies was constructed by the authors alone, and is therefore far from definitive, but provides a useful illustration of the kind of output this methodology can produce.

Based on this causal map, we can identify three candidates for canary capabilities:

- Representations that allow variable-binding and disentanglement: the ability to construct abstract, discrete and disentangled representations of inputs, to allow for efficiency and variable-binding. We hypothesise that this capability underpins several others, including grammar, mathematical reasoning, concept formation, and flexible memory.
- Flexible memory: the ability to store, recognise, and re-use memory and knowledge representations. We hypothesise that this ability would unlock many others, including the ability to learn from dynamic data, to learn in a continual fashion, and to update old interpretations of data as new information is acquired.
- Positing unobservables: the ability to recognise and use unobservable concepts that are not represented in the visual features of a scene, including numerosity or intentionality.

We might tentatively suggest that these are important capabilities to track progress on from the perspective of anticipating HLMI.

6. Discussion and Future Directions

As the two illustrative examples show, there are many complexities and challenges involved in putting this method into practice. One particular challenge is that there is likely to be substantial uncertainty in the causal graphs developed. This uncertainty can come in many forms.

Milestones that are not well understood are likely to be composed of several sub-milestones. As more research is produced, the graph will be in need of revision. Some such revisions may include the addition of connections between milestones that were previously not foreseen, which in turn might alter the number of outgoing connections from nodes and turn them into potent nodes, i.e. "canaries".

The process of involving a diversity of experts in a multi-stage, collaborative process is designed to reduce this uncertainty by allowing for the identification of nodes and relationships that are widely agreed upon and so more likely to be robust. However, considerable uncertainty will inevitably remain due to the nature of forecasting. The higher the level of abstraction and ambiguity in the events studied (like events such as HLMI, which we use for our illustration) the greater the uncertainty inherent in the map and the less reliable the forecasts will likely be. It will be important to find ways to acknowledge and represent this uncertainty in the maps developed and conclusions drawn from them. This might include marking uncertainties in the graph and taking this into account when identifying and communicating "canary" nodes.

Given the uncertainty inherent in forecasting, we must consider what kinds of inevitable misjudgements are most important to try to avoid. A precautionary perspective would suggest it is better to slightly overspend resources on monitoring canaries that turn out to be false positives, rather than to miss an opportunity to anticipate significant technological impacts. This suggests we may want to set a low threshold for what should be considered a "canary" in the final stage of the method.

The uncertainty raises an important question: will it on average be better to have an imperfect, uncertain mapping of milestones rather than none at all? There is some chance that incorrect estimates of "canaries" could be harmful. An incorrect mapping could focus undue attention on some avenue of AI progress, waste resources or distract from more important issues.

Our view is that it is nonetheless preferable to attempt a prioritisation. The realistic alternative is that anticipatory governance is not attempted or informed by scholars' individual estimates in an ad-hoc manner, which we should expect to be incorrect more often than our collective and structured expert elicitation. How accurate our method is can only be studied by trialling it and tracking its predictions as AI research progresses to confirm or refute the forecasts.

Future studies are likely to face several trade-offs in managing the uncertainty. For example, a large and cognitively diverse expert group may be better placed to develop robust maps eventually, but this may be a much more challenging process than doing it with a smaller, less diverse group — making the latter a tempting choice (see Rask for a discussion of this trade-off).[84] The study of broad and high-level questions (such as when we might attain HLMI or automate a large percentage of jobs) may be more societally relevant or intellectually motivating, but narrower studies focused on nearer-term, well-defined applications or impacts may be easier to reach certainty on.

A further risk is that this method, intended to identify warning signs so as to give time to debate transformative applications, may inadvertently speed up progress towards AI capabilities and applications. By fostering expert deliberation and mapping milestones, it is likely that important research projects and goals are highlighted and the field's research roadmap is improved. This means our method must be used with caution.

However, we do not believe this is a reason to abandon the approach, since these concerns must be balanced against the benefits of being able to deliberate upon and shape the impacts of AI in advance. In particular, we believe that the process of distilling information from experts in a way that can be communicated to wider society, including those currently underrepresented in debates about the future of AI, is likely to have many more benefits than costs.

The idea that we can identify "warning signs" for progress assumes that there will be some time lag between progress on milestones, during which anticipatory governance work can take place. Of course, the extent to which this is possible will vary, and in some cases, unlocking a "canary" capability could lead to very rapid progress on subsequent milestones. Future work could consider how to incorporate assessment of timescales into the causal graphs developed, so that it is easier to identify canaries which warn of future progress while allowing time to prepare.

Future work should also critically consider what constitutes relevant "expertise" for the task of identifying canaries, and further explore ways to effectively integrate expert knowledge with the values and perspectives of diverse publics. Our method finds a role for the expert situated in a larger democratic process of anticipating and regulating emerging technologies. Expert judgement can thereby be beneficial to wider participation. However, processes that allow more interaction between experts and citizens could be even more effective. One limitation of the method presented in this chapter is that it requires one to have already identified a particular transformative event of concern, but does not provide guidance on how to identify and prioritise between events. It may be valuable to consider how citizens that are impacted by technology can play a role in identifying initial areas of concern, which can then feed into this process of expert elicitation to address the concerns.

7. Conclusion

We have presented a flexible method for identifying early warning signs, or "canaries" in AI progress. Once identified, these canaries can provide focal points for anticipatory governance efforts, and can form the basis for meaningful participatory processes enabling citizens to steer AI developments and their impacts. Future work must now test this method by putting it into practice, which will more clearly reveal both benefits and limitations. Our artificial canaries offer a chance for forward-looking, democratic assessments of transformative technologies.

Acknowledgements

We thank reviewers for their particularly detailed comments and engagement with this chapter, the scholars at the Leverhulme Centre for the Future of Intelligence for fruitful discussions after our presentation, as well as the attendees of the workshop "Evaluating Progress in AI" at the European Conference on AI (August 2020) for recognising the potential of this work. We particularly thank Carolyn Ashurst and Luke Kemp for their efforts and commentary on our drafts.

Appendix available online at https://doi.org/10.11647/OBP.0360#resources

Notes and References

[1] Crawford, K. et al. 'AI Now Report 2019', *AI 2019 Report* (2019), p. 100.

[2] Russell, S. *Human Compatible*. Viking Press (2019); Cath, C., S. Wachter, B. Mittelstadt, M. Taddeo and L. Floridi. 'Artificial Intelligence and the "good society": The US, EU, and UK approach', *Sci. Eng. Ethics, 24*(2) (April 2018): 505–28. https://doi.org/10.1007/s11948017-9901-7. Whittlestone, J., R. Nyrup, A. Alexandrova, K. Dihal and S. Cave. *Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: A Roadmap for Research* (2019), p. 59. Dwivedi, Y. K. et al. 'Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy', *Int. J. Inf. Manag.* (August 2019), p. 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

[3] Gruetzemacher, R. and J. Whittlestone. 'The transformative potential of Artificial Intelligence', *ArXiv191200747 Cs* (September 2020). http://arxiv.org/abs/1912.00747

[4] Brundage, M. et al., 'The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation', *ArXiv180207228 Cs* (February 2018). http://arxiv.org/abs/1802.07228

[5] Howard, P. *Lie Machines, How to Save Democracy From Troll Armies, Deceitful Robots, Junk News Operations, and Political Operatives*. Yale University Press (2020).

[6] Frey, C. B. and M. A. Osborne. 'The future of employment: How susceptible are jobs to computerisation?', *Technol. Forecast. Soc. Change, 114* (January 2017): 254–80. https://doi.org/10.1016/j.techfore.2016.08.019

[7] Grace, J. K., J. Salvatier, A. Dafoe, B. Zhang and O. Evans. 'Viewpoint: When will AI exceed human performance? Evidence from AI experts', *Artif. Intell. Res., 62* (July 2018): 729–54. https://doi.org/10.1613/jair.1.11222; Cremer, C. Z. 'Deep limitations? Examining expert disagreement over deep learning', *Prog. Artif. Intell.* Springer (to be published 2021).

[8] Collingridge, D. *The Social Control of Technology*. Frances Pinter (1980).

[9] Etzioni, O. 'How to know if Artificial Intelligence is about to destroy civilization', *MIT Technology Review*. https://www.technologyreview.com/s/615264/artificial-intelligence-destroy-civilization-canaries-robotoverlords-take-over-world-ai/; Dafoe, A. 'The academics preparing for the possibility that AI will destabilise global politics', *80,000 Hours* (2018). https://80000hours.org/ podcast/episodes/allan-dafoe-politics-of-ai/

[10] Grace et al. (2018); Müller, V. C. and N. Bostrom. 'Future progress in Artificial Intelligence: A survey of expert opinion', in *Fundamental Issues of Artificial Intelligence*, ed. V. C. Müller. Springer (2016), pp. 555–72; Baum, S. D., B. Goertzel and T. G. Goertzel. 'How long until human-level AI? Results from an expert assessment', *Technol. Forecast. Soc. Change, 78*(1) (January 2011): 185–95. https://doi.org/10.1016/j.techfore.2010.09.006; Beard, S., T. Rowe and J. Fox, 'An analysis and evaluation of methods currently used to quantify the likelihood of existential hazards', *Futures, 115* (January 2020), p. 102469. https://doi.org/10.1016/j.futures.2019.102469

[11] Baum et al. (2011).

[12] Müller and Bostrom (2016).

[13] Grace et al. (2018).

[14] Cremer (2021); Tetlock, P. E. and D. Gardner. *Superforecasting: The Art and Science of Prediction* (1st edition). Crown Publishers (2015).

15 Grace et al. (2018).

16 E.g. Benaich, N. and I. Hogarth. *State of AI Report 2020* (2020). https://www. stateof. ai/; Eckersley, P. and Y. Nasser. 'AI progress measurement', *Electronic Frontier Foundation* (12 June 2017). https://www.eff.org/ai/metrics; 'Papers with code'. https://paperswithcode.com; Perrault, R. et al. 'The AI Index 2019 annual report', *AI Index Steer. Comm. Hum.-Centered AI Inst. Stanf. Univ. Stanf.* (2019).

17 Gruetzemacher. 'A holistic framework for forecasting transformative AI', *Big Data Cogn. Comput., 3*(3) (June 2019): 35. https://doi.org/10.3390/bdcc3030035

18 Linstone, H. A. and M. Turoff. *The Delphi Method*. Addison-Wesley Reading (1975).

19 West, S. M., M. Whittaker and K. Crawford. 'Discriminating systems: Gender, race and power in AI', *AI Now Institute* (2019). https://ainowinstitute.org/discriminatingsystems.html

20 Nemitz, P. and M. Pfeffer. *Prinzip Mensch — Macht, Freiheit und Demokratie im Zeitalter der Künstlichen Intelligenz*. Verlag J.H.W. Dietz Nachf. (2020).

21 Ipsos, M. 'Public views of Machine Learning: Findings from public research and engagement conducted on behalf of the Royal Society', The Royal Society (2017). https://royalsociety.org/-/ media/policy/projects/machine-learning/publications/public-views-ofmachine-learning-ipsos-mori.pdf; The RSA. 'Artificial Intelligence: Real public engagement', *Royal Society for the Encouragement of Arts, Manufactures and Commerce* (2018); Cohen, T., J. Stilgoe and C. Cavoli. 'Reframing the governance of automotive automation: insights from UK stakeholder workshops', *J. Responsible Innov., 5*(3) (September 2018): 257–79. https://doi.org/10.1080/23299460.2018.1495030

22 Lengwiler, M. 'Participatory approaches in science and technology: Historical origins and current practices in critical perspective', *Sci. Technol. Hum. Values, 33*(2) (March 2008): 186–200. https://doi.org/10.1177/0162243907311262; Rask, M. 'The tragedy of citizen deliberation — Two cases of participatory technology assessment', *Technol. Anal. Strateg. Manag., 25*(1) (January 2013): 39–55. https://doi.org/10.1080/09537325.2012.751012

23 Chilvers, J. 'Deliberating competence: Theoretical and practitioner perspectives on effective participatory appraisal practice', *Sci. Technol. Hum. Values, 33*(2) (March 2008): 155–85. https://doi.org/10.1177/0162243907307594

24 Ipsos (2017).

25 Abels, G. 'Participatory technology assessment and the "institutional void": Investigating democratic theory and representative politics', in *Democratic Transgressions of Law* (vol. 112). Brill (2010), pp. 237–68.

26 Abels (2010).

27 Biegelbauer, P. and A. Loeber. 'The challenge of citizen participation to democracy', *Inst. Für Höhere Stud. — Inst. Adv. Stud. IHS* (2010), p. 46.

28 Rowe, G. and L. J. Frewer. 'A typology of public engagement mechanisms', *Sci. Technol. Hum. Values, 30*(2) (April 2005): 251–90. https://doi.org/10.1177/0162243904271724

29 Abels (2010).

30 Biegelbauer and Loeber (2010).

31 Rask (2013).

32 Hong, L. and S. E. Page. 'Groups of diverse problem solvers can outperform groups of high-ability problem solvers', *Proc. Natl. Acad. Sci., 101*(46) (November 2004):

16385–89. https://doi.org/10.1073/pnas.0403723101; Landemore, H. *Democratic Reason*. Princeton University Press (2017).

33 Joss, S. and S. Bellucci. *Participatory Technology Assessment: European Perspectives*. Center for the Study of Democracy (2002); Zhao, Y., C. Fautz, L. Hennen, K. R. Srinivas and Q. Li. 'Public engagement in the governance of science and technology', in *Science and Technology Governance and Ethics: A Global Perspective From Europe, India and China*, ed. M. Ladikas, S. Chaturvedi, Y. Zhao and D. Stemerding. Springer International Publishing (2015), pp. 39–51; Rask, M. T. et al. *Public Participation, Science and Society: Tools for Dynamic and Responsible Governance of Research and Innovation*. Routledge — Taylor & Francis Group (2018).

34 Burgess, J. and J. Chilvers. 'Upping the ante: a conceptual framework for designing and evaluating participatory technology assessments', *Sci. Public Policy, 33*(10) (December 2006): 713–28. https://doi.org/10.3152/147154306781778551

35 Rask (2013); Abels (2010).

36 Hsiao, Y. T., S.-Y. Lin, A. Tang, D. Narayanan and C. Sarahe. 'vTaiwan: An empirical study of open consultation process in Taiwan', *SocArXiv* (July 2018). https://doi.org/10.31235/osf.io/xyhft

37 Rask et al. (2018).

38 Joss and Bellucci (2002).

39 Zhao et al. (2015); Hansen, J. 'Operationalising the public in participatory technology assessment: A framework for comparison applied to three cases', *Sci. Public Policy, 33*(8) (October 2006): 571–84. https://doi.org/10.3152/147154306781778678

40 Burgess and Chilvers (2006).

41 Ertiö, T.-P., P. Tuominen and M. Rask. 'Turning ideas into proposals: A case for blended participation during the participatory budgeting trial in Helsinki', in *Electronic Participation: ePart 2019* (Jul. 2019), pp. 15–25. https://doi.org/10.1007/978-3-030-27397-2_2

42 Rowe Frewer (2005).

43 Rask, M. 'Foresight — Balancing between increasing variety and productive convergence', *Technol. Forecast. Soc. Change — TECHNOL FORECAST SOC CHANGE, 75* (October 2008): 1157–75. https://doi.org/10.1016/j.techfore.2007.12.002

44 Mauksch, S., H. A. von der Gracht and T. J. Gordon. 'Who is an expert for foresight? A review of identification methods', *Technol. Forecast. Soc. Change, 154* (May 2020), p. 119982. https://doi.org/10.1016/j.techfore.2020.119982

45 Lengwiler (2008).

46 Chilvers (2008).

47 Saldivar, J., C. Parra, M. Alcaraz, R. Arteta and L. Cernuzzi. 'Civic technology for social innovation: A systematic literature review', *Comput. Support. Coop. Work CSCW, 28*(1–2) (April 2019): 169–207. https://doi.org/10.1007/s10606-018-9311-7

48 Kariotis, T. and J. Darakhshan. 'Fighting back algocracy: The need for new participatory approaches to technology assessment', in *Proceedings of the 16th Participatory Design Conference 2020 — Participation(s) Otherwise — Volume 2*. Manizales Colombia (June 2020), pp. 148–53. https://doi.org/10.1145/3384772.3385151; Whitman, M., C. Hsiang and K. Roark. 'Potential for participatory big data ethics and algorithm design: A scoping mapping review', *Proceedings of the 15th Participatory Design Conference: Short*

Papers, Situated Actions, Workshops and Tutoriall — Volume 2 (August 2018), pp. 1–6. https://doi.org/10.1145/3210604.3210644

49 Buckner, C. and K. Yang. 'Mating dances and the evolution of language: What's the next step?', *Biol. Philos., 32* (2017). https://doi.org/10.1007/s10539017-9605-z

50 LeCun, Y., Y. Bengio and G. Hinton. 'Deep learning', *Nature, 521*(7553) (May 2015): 436–44. https://doi.org/10.1038/nature14539

51 Mauksch et al. (2020).

52 Landemore (2017); Page, S. E. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools and Societies*. Princeton University Press (2008).

53 Scavarda, A. J., T. Bouzdine-Chameeva, S. M. Goldstein, J. M. Hays and A. V. Hill. 'A review of the causal mapping practice and research literature', in *Abstract Number: 002-0256* (2004), p. 21.

54 Scavarda et al. (2004).

55 Markíczy, L. and J. Goldberg. 'A method for eliciting and comparing causal maps', *J. Manag., 21*(2) (January 1995): 305–33. https://doi.org/10.1016/0149-2063(95)90060-8; Eden, C. and F. Ackermann. 'Cognitive mapping expert views for policy analysis in the public sector', *Eur. J. Oper. Res., 152*(3) (February 2004): 615–30. https://doi.org/10.1016/S0377-2217(03)00061-4

56 Eden, C. *On the Nature of Cognitive Maps* (1992). https://doi.org/10.1111/ J.1467-6486.1992.TB00664.X

57 Scavarda et al. (2004).

58 Ackerman, F., J. Bryson and C. Eden. *Visible Thinking, Unlocking Causal Mapping for Practical Business Results*. John Wiley & Sons (2004).

59 Montibeller, G. and V. Belton. 'Causal maps and the evaluation of decision options — A review', *J. Oper. Res. Soc., 57*(7) (July 2006): 779–91. https://doi.org/10.1057/palgrave.jors.2602214

60 Scavarda, A. J., T. Bouzdine-Chameeva, S. M. Goldstein, J. M. Hays and A. V. Hill. 'A methodology for constructing collective causal maps*', *Decis. Sci., 37*(2) (May 2006): 263–83. https://doi.org/10.1111/j.15405915.2006.00124.x

61 Eden (1992).

62 Eden, C., F. Ackermann and S. Cropper. 'The analysis of cause maps', *J. Manag. Stud., 29*(3) (1992): 309–24. https://doi.org/10.1111/j.1467-6486.1992.tb00667.x

63 Ackerman, Bryson and Eden (2004).

64 Scavardia et al. (2004); Scavardia et al. (2006).

65 Eden and Ackermann (2004).

66 Ackermann, F.and C. Eden. 'Using causal mapping with group support systems to elicit an understanding of failure in complex projects: Some implications for organizational research', *Group Decis. Negot., 14*(5) (September 2005): 355–76. https://doi.org/10.1007/s10726-005-8917-6

67 Eden, C. F. Ackermann, J. Bryson, G. Richardson, D. Andersen and C. Finn. *Integrating Modes of Policy Analysis and Strategic Management Practice: Requisite Elements and Dilemmas* (2009), p. 13.

68 Hsiao et al. (2018).

69 Neudert, L.-M. and P. Howard. 'Ready to vote: Elections, technology and political campaigning in the United Kingdom', *Oxford Technology and Elections Commission* (October 2019). https://apo.org.au/node/263976; Bolsover, G. and P. Howard. 'Computational propaganda and political big data: moving toward a more critical research agenda', *Big Data, 5*(4) (December 2017): 273–76. https://doi.org/czz; Mazarr, M. J., R. Bauer, A. Casey, S. Heintz and L. J. Matthews. *The Emerging Risk of Virtual Societal Warfare: Social Manipulation in a Changing Information Environment* (October 2019). https://www.rand.org/pubs/research_reports/ RR2714.html

70 Wu, T. *The Attention Merchants: From the Daily Newspaper to Social Media, How Our Time and Attention is Harvested and Sold*. Atlantic Books (2017).

71 Howard (2020); Starbird, K. 'Disinformation's spread: bots, trolls and all of us', *Nature, 571*(7766) (July 2019): 449–50.

72 Like R. Gorwa and D. Guilbeault. 'Unpacking the social media bot: A typology to guide research and policy', *Policy Internet, 12*(2) (June 2020): 225–48. https://doi.org/10.1002/poi3.184

73 Ferrara, E. 'Disinformation and social bot operations in the run up to the 2017 French Presidential Election', *Social Science Research Network* (June 2017). https://doi.org/10.2139/ ssrn.2995809

74 Shao, C., G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini and F. Menczer. 'The spread of low-credibility content by social bots', *Nat. Commun.*, *9*(1) (November 2018). https://doi.org/10.1038/s41467-01806930-7

75 Howard, P. N., S. Woolley and R. Calo. 'Algorithms, bots, and political communication in the US 2016 election: The challenge of automated political communication for election law and administration', *J. Inf. Technol. Polit., 15*(2) (April 2018): 81–93. https://doi.org/10.1080/19331681.2018.1448735

76 Chessen, M. 'The MADCOM future: How Artificial Intelligence will enhance computational propaganda, reprogram human culture, and threaten democracy... and what can be done about it', *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC Press (2018), pp. 127–44.

77 Kertysova, K. *Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered* (2018). https://doi.org/10.1163/18750230-02901005

78 Brainard, J. and P. R. Hunter. 'Misinformation making a disease outbreak worse: Outcomes compared for influenza, monkeypox, and norovirus', *SIMULATION, 96*(4) (April 2020): 365–74. https://doi.org/10.1177/0037549719885021; Seger, E., S. Avin, G. Pearson, M. Briers, S. O Heigeartaigh and H. Bacon. *Tackling Threats to Informed Decision-Making in Democratic Societies: Promoting Epistemic Security in a Technologically Advanced World*. Allan Turing Institute, CSER (2020). https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

79 Jamieson, K. H. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President: What We Don't, Can't, and Do Know*. Oxford University Press (2020).

80 Howard (2020).

81 Howard (2020).

82 Grace et al. (2018); Müller and Bostrom (2016).

83 Cremer (2021).

84 Rask (2008).

IV. POLICY, INSTITUTIONS, AND IMPACTS

The chapters in this, the final section, of the collection are drawn together by their shared concern with the question of "what is to be done?". While every chapter in this volume has shared the desire to understand extreme global risk as a means of reducing it, these chapters focus on the policies, institutions, and processes that are needed to achieve this goal. While these chapters are markedly different in terms of their objects of concern (covering everything from national institutions and policy to global diplomacy and institutional investors), the heuristics with which they understand the possibility of extreme global risk and why we should care about it (ranging from "internal" institutional logics to abstract ethical ideals), and indeed the nature of the proposals proffered as to what might be done about them (including specific policies and institutions and more general proposals, frameworks, and research agendas), we can still usefully trace similarities and common themes across each of them.

Foremost among these is their direct interest in shaping actual policies, institutional behaviours, and governance priorities in the real world. There is a long tradition in existential risk research, dating back to the work of people like Bertrand Russell and H. G. Wells, to believe that extreme global risks demand ideal solutions such as "world government", total surveillance, or revolutions in human behaviour. However, these chapters are all solidly grounded in the realities of the 21st century policy landscape. In Chapter 23, *Financing Our Final Hour*, for example, the authors build an impressive and empirically robust case for the urgent necessity for institutional investors to take seriously their responsibilities to the people and planet that their profits are predicated upon by adhering to a Financial Hippocratic Oath. Chapter 21, *It Takes a Village: The Shared Responsibility of Raising an Autonomous Weapon*, by contrast, presents a study of how to translate notions of *shared* responsibility and the embedding of strong norms into the back-and-forth of defence policy, technological design, and military procurement. Despite the varied focus of the chapters in this section, they each present an approach to deploying empirically robust, intellectually ambitious, and contextually sensitive research into policy proposals, engagement tools, and other forms of policy engagement.

The types of policy work undertaken across these chapters is far from uniform. The proposals developed in the chapters range from the pragmatic to the speculative, and the approaches advocated by the authors range from emphasising the requirement for evidence and informed decision-making through to means of fostering dialogue and embracing uncertainty.

Another common theme among the chapters gathered here is that they are concerned — in one way or another — with questions that go beyond issues of necessity and duress (that is, what simply *must* be done if we are to survive at all) and also look towards how social, political, and economic conditions more amenable to human and planetary survival might be fostered. Again, the chapters take a variety of approaches to exploring these questions. Chapter 22 examines the ways that future generations might be best represented in the policy-making process of the present, while in Chapter 20, Paul Ingram's account of fostering dialogue and acceptance in the fraught and (perhaps necessarily) adversarial world of nuclear disarmament diplomacy prompts us to directly consider the ways that work in the field can open up spaces where the politics of possibility can come more clearly into view. That is not to say that these chapters are idealistic or utopian, although they sketch possibilities of existential hope, futures where better decision making at many levels can both safeguard us from catastrophic futures and guide us towards better ones. However, in contrast to some earlier work in the field of Existential Risk Studies, they do not make any claims about what kind of future would be best for humanity, and indeed make suggestions that would bring human futures under greater democratic control, leaving this as a question that individuals are left to answer for themselves.

Still, imagining what *might* be possible is central to any work that talks about the conditions of the future — whether its ostensible concern is the mitigation of risk, or the fostering of a utopia. The contributions in this section are perhaps conspicuous in this sense, primarily for the directness with which they explore these political, ontological, and normative relationships between our present and our possible futures. In asking how a researcher or policy-maker might contribute towards the attainment of a safer or more survivable future through the development of policy or through influencing the trajectory or structure of institutions, the authors each draw us towards important foundational assumptions related to both how we understand the world *as it is*, and how we imagine *it ought to be*. In this way, whilst also providing rich accounts of how policies and institutions might evolve in response to research and understanding of specific or systemic catastrophic risks, the chapters also provide an opportunity to reflect on some of the underlying factors that condition work in the field — with a range of perspectives on political, social, geographic and temporal relationalities presented by the authors. Thus, Chapter 19, *The Cartography of Global Catastrophic Governance*, charts the concentrations of different governance efforts related to catastrophic risks areas and argues for both greater attention and greater coordination, while Chapter 18, *Pathways to Linking Science and Policy in Global Risk*, provides ideas for researchers to identify pathways to impact for their own work.

The section opens with two chapters that provide a general summary of the current state of policy around extreme global risk. Chapter 18 is based on an assessment of six policy engagement activities at CSER, reflecting on the meaning of impactful research and highlighting the different ways in which this can be achieved. The chapter provides a call to arms for all researchers in this space to embed impact within their research but also highlights the importance of setting clear goals for this activity and of practicing continuous evaluation to ensure that these are being met.

Chapter 19 charts the efficacy and concentration of different GCR governance efforts, proposing a typology that allows for comparisons based on risk focus, institutional arrangement, and effectiveness of implementation. The chapter draws attention to those areas that have received more significant attention, and those which have thus far been under-attended to as either drivers of catastrophic risk or as factors that determine the vulnerability of a society to GCRs. Overall, the authors argue that several GCR hazards (climate change, nuclear weapons) are covered by international law but often inadequately. They note that institutions often lack clear enforcement and compliance mechanisms, or have been unable to address the underlying collective action problem.

Other issues, such as solar geoengineering, catastrophic uses of AI, some areas of ecological collapse, the chapter argues to be relatively neglected. Just as understanding the drivers and catalyst of catastrophic events is complex, so too is understanding the differential landscape of efforts to govern these risk areas. The landscape across GCR governance is fragmented and mandates both within and across different hazards and vulnerabilities. The authors do note, however, that there are a number of approaches that can be taken to enhance understanding of GCRs and to increase resilience to them even in the face of a high degree of uncertainty.

The remaining chapters focus on specific proposals for reducing extreme global risk within particular contexts and targeted at different actors. In Chapter 20, Paul Ingram provides an account that sheds light on the ways that the challenges of developing fit-for-purpose policy (which were highlighted at a macro-level in the taxonomy of Chapter 19) can play out at the level of international and interpersonal relationships within the context of nuclear diplomacy. Ingram's reflexive account draws our attention not only to the geopolitical competitions and power differentials that can frustrate idealised imaginaries of global cooperation to reduce the potential for global disasters, but also reminds us that we cannot — in our efforts to understand problems of global scale — leave out the realities, constraints, and possibilities that are created and perpetually re-negotiated by both people and states. Indeed, while Ingram's aim in this chapter is to explicate his "Stepping Stones" approach, and to provide some generalisable insights into how open dialogue and iterative processes embrace the potentiality created by even the most modest of incremental change, it is also a welcome re-insertion of the international into our discussion of how to think through the global. Ingram's close attention to navigating relationality, and his central drive to open up spaces of possibility, makes his contribution to this volume a provocative and pedagogically engaging one. If there is a core message conveyed by his chapter, it is a straightforward but powerful one: just because things are as they are, we should not assume that they need to be that way, nor that they cannot be changed.

Several of the chapters here rely heavily on the appeal to a central metaphor. For Kemp and Rhodes in Chapter 19, it is the mapping of a landscape, for Ingram, it is the image of the act of crossing an obstacle (and perhaps meeting in the middle) using stepping stones. In Chapter 21, Avin and Jayanata rely instead on an aphoristic analogy, creatively deploying a communitarian imaginary of child nurturing to frame their discussion of LAWS governance. Whereas Kemp and Rhodes examine the difference between different extant approaches to GCR governance, and Ingram reflects on the layered inter-relationality between individuals and between those with shared and competing visions, Avin and Jayanata look instead towards the relationships between different elements of a complex sociotechnical system: the development, use, and regulation of Lethal Autonomous Weapons. The chapter recounts the results and insights garnered over the course of a series of interviews with experts based across the UK, that were framed as a mock parliamentary inquiry. The key findings concerned a lack of accountability and the malleability of concepts such as "meaningful human control" that can become unhelpfully restrictive and reductive if principles of collective responsibility are not embedded in the process of creating, procuring, training with, and using autonomous weapons.

Their contribution is instructive in demonstrating the sometimes less visible contingency and specificity of institutional and regulatory arrangements that bely the technological artifacts that are the focus of risk research and governance. When we consider, for example, the potentialities of an autonomous weapon or a neural network, we are in fact considering a far larger, more diffuse and complex web of people, institutions, manufacturers, rules, and norms that must be considered within a richer account of a technology and its effects.

Making sense of catastrophe is difficult. Contributions throughout the volume, show us that some scholars have approached this difficulty by embracing truly *global* approach to extreme global risks — understanding risks within a framework that engages explicitly with both the planet, and humanity as totalities where hazards and vulnerabilities must be measured against the world, its population and its future at the level of their largest aggregations. The chapters in this section explore some of the ways that both ethical imperatives — concerning generational inequity and representation — and observable complexities (such as the interrelationships between, and concentrations of governance within, different risk areas and the asymmetric distribution of vulnerabilities to them) can be illuminated and further explored through a variety of different case-studies in policy development or institutional change. In Chapter 22, for example, Natalie Jones explores the contours of an argument concerning the ethical imperative to consider the equal standing, and rights of, future generations. The case might initially appear reminiscent of calls from the utilitarian aspects of what earlier contributions to this volume labelled the "TUA" — yet Jones' treatment of the issue as one that troubles the boundaries of democratic norms, and which requires ethical sensitivity rather than moral doctrine, leads her to produce a work that is both pragmatic in its development of recommendations yet nuanced in its treatment of what is generalisable and what is contextually specific.

Jones examines the development of future generations policies in a number of states, including Scotland, Israel, Finland and Hungary. The chapter assesses the strengths and weaknesses of different types of institutional arrangements for the representation of future people in present day governance processes, whilst further developing a set of recommendations on the basis of both an assessment of the UK specific context and the comparative analysis of existing policies. Foremost among these is the recommendation for creating a APPG for Future Generations — a recommendation that was taken up by Parliament in 2017 due to the direct efforts of the chapter authors with the support of CSER.

The final chapter in this section, and indeed in the collection as a whole, is *Financing Our Final Hour* by Kemp, Belfield, Quigley, Weitzdörfer, and Beard. This transitions from a focus on governments and international bodies to considering private sector responsibilities and, in particular, the role of large institutional investors. The chapter was developed over a number of years as a collective response to challenging moral, political, and economic discussions around the global divestment from the Fossil Fuels campaign, alternative approaches for financial institutions to tackle climate change, and the relevance of these to other sources of extreme global risk. This work helped to influence the University of Cambridge in its decision to divest from fossil fuels in 2021, with Dr Quigley taking a secondment to the university's Chief Financial Officer to help develop this policy. However, this chapter presents the totality of this work for the first time. The chapter considers both the (ethical, legal, financial, and prudential) reasons why large institutional investors should care about global risk and the different tactics available to them to achieve these aims. Ultimately it concludes that all institutional investors have reason to adopt policies to stop contributing to extreme global risk, but that the very largest, so-called "universal owners" that represent an entire slice of the global economy should go beyond this and actively use their investments to reduce global risk by all available means.

It should of course go without saying that the policies and proposals presented in this section provide only a small subset of the many actions that are required to achieve the goal of reducing the level of extreme global risk, as well as the many more targeted proposals designed to tackle climate change, biosecurity threats, nuclear war, natural global-scale disasters, and the responsible innovation and deployment of new technologies. We have selected these chapters, and the chapters in this volume more generally, to be representative of a diversity of ways of thinking, united in their concern to understand and reduce extreme global risk and their commitment to contribute to an open, pluralistic, transparent, and robust field of Existential Risk Studies, but divergent in their community of stakeholders, method of construction, and locus of concern. Ultimately our hope is not that anyone should read these chapters and know what needs to be done to reduce extreme global risk, but rather that they will read these chapters and gain a better understanding of the gaps that must be filled in order to achieve this aim, and a confidence that they too can play a role in filling these.

18. Pathways to Linking Science and Policy in Global Risk

Clarissa Rios Rojas, Catherine Richards, Catherine Rhodes and Paul Ingram

Highlights:

- Existential Risk Studies is an action-oriented discipline that must embed policy impact to reduce global risk as an essential part of our work.
- Impact comes in many forms, and it is good to plan for impact and engage with relevant stakeholders from the earliest stages of the research process.
- CSER has engaged in a variety of impact-focused activities and has found our broad network, high-quality research, and willingness to shape our work to fit the needs of policy-makers essential to our success.
- It is important to set clear goals for impact and continuously monitor the policy landscape and how it is changing.

This chapter is based on a CSER report published in July 2021, which is a practical how-to guide to engaging with stakeholders and policy-makers. It involved one-to-one interviews and a workshop designed to elicit advice and experience from CSER researchers. After presenting a general overview of CSER's approach to impact, the chapter outlines six policy case studies to develop advice for others seeking to influence policy. The "how to" character of this chapter is usefully complemented by the macro-analyses of policy-shaping contained in Chapter 19, and is

https://doi.org/10.11647/OBP.0360.18

further supplemented by the reflective account of seeking to influence nuclear disarmament diplomacy in Chapter 20.

When studying existential and Global Catastrophic Risk, we also look for how to manage and reduce it. This goes beyond more traditional academic outputs and leads to possible engagement with policy-makers and other stakeholders that influence the systems that generate or mitigate risk. The Centre for the Study of Existential Risk (CSER) looks to strengthen the impact of our research on practical policy and has developed and improved our methods in doing so. In this chapter, we look at effective approaches and relevant skills to promote impact (such as project management, communication, networking, expertise and familiarity with the policy landscape), and suggest step-by-step guidance to assist in planning interventions.

1. Policy Impact

Academic impact refers to influence within the academic community. This can be demonstrated, for example, by shifting old dogmas or by contributing to new theories across and within disciplines. Policy impact, by contrast, refers to contributions with social, economic and political dimensions. This includes diverse reference groups and transitions, such as technological progress, government regulation, or corporate management.[1]

Working to achieve policy impact is a pathway to applying scientific evidence to achieve a better world. Doing so can expand your published materials by turning the experience into academic papers or your work features in official documents. It can improve your network amongst academics and others looking for impact, triggering a virtuous circle of new contacts, research, collaborations, and ideas for further impact. It allows academics to transition from abstract to applied practical work. It can also bring in extra funding — as funders and research councils are increasingly impact-focused — as well as opening opportunities for consultancy or follow-on careers, and increasing your institution's reputation.

Policy engagement is often time-consuming and may distract from other priorities (such as scientific publications), which may have a clearer linkage with academic career progression. It also involves a significant risk of uncertain or minimal success that is difficult to track or credit.

Researchers interested in increasing their policy impact can work with stakeholders in three sectors: civil society,[2] government[3] and business.[4] The Cambridge Public Policy report "How to Evidence and Record Policy Impact"[5] explores impacts on UK public policy and provides indicators that researchers and institutions can use to evaluate the influence of their research in this sphere. These include, amongst others: citations in government reports or international bodies; changing public understanding of a policy issue or challenge; engagement with campaign and pressure groups, and other civil society organisations; and improving public services. These indicators are based on the Research Excellence Framework (REF) process, which is used to assess research performance at academic institutions in the UK. CSER provided an impact case study for REF2021, which starts by observing that:

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

2. Approaches to Policy Engagement

Academics and think tanks often make the mistake of seeing policy engagement as an afterthought once the research product (report or article) has been published. It is better to incorporate impact considerations right at the start when considering purpose and study design, which also creates opportunities for policy co-creation. Including stakeholders in expert solicitation processes and other participatory research methods can also be effective. It is perfectly consistent with integrity to be asking what the priorities of target policy "customers" might be and how to develop the research questions that shape your research, even if that might be to challenge established practices head-on.

A CSER policy engagement might start by contacting science-policy brokers, such as the United Nation's various offices (WHO, UNDRR, UNODA, etc), the International Science Council, the Centre for Science and Policy (CSAP) at the University of Cambridge, the Royal Society, the Simon Institute for Longterm Governance, or the Centre for Long Term Resilience (CLTR), among others. We also monitor official and parliamentary websites for relevant opportunities to contribute to policy enquiries and consultations. These could include calls for papers or open consultations from the UN-affiliated bodies or the European Commission. It can also be effective to join relevant expert advisory groups within governments, inter-governmental and non-governmental organisations. It may be possible to join scientific networks — such as the Global Young Academy or the International Network for Science Government Advice — national-level science academies, or engage with social movements.

A common approach would be to write a policy paper. This can be self-published or placed in a relevant journal, such as the Cambridge Journal of Science and Policy, or on industry or think-tank websites, newsletters, or blogs. There may be opportunities to work more directly with official organs. For example, the UK's Parliamentary Office on Science and Technology publishes "notes", four-page briefings that review emerging research areas.

3. Six CSER Pathways

We used six study cases of CSER's policy impact to develop advice for others seeking to influence policy:

3.1 Creating an All Party Parliamentary Group (APPG) on future generations

This began as a student initiative with CSER advising research students at the University of Cambridge, who were looking at different ways in which future generations were represented in politics around the world. This led to a paper published in the journal *Futures* and recommended establishing a group to represent future generations in the UK parliament.[6] CSER further collaborated with students to identify and engage with key people, including parliamentarians and people with experience in setting up such groups, to understand how such a group could be established. Engaging a broad range of key people to widen political buy-in was key. Once enough support was obtained, it was possible to complete the UK Parliament's standard template to form APPGs and approve it with a sufficiently large and broad group of parliamentarians.

3.2 AI white paper for House of Lords

This opportunity emerged following an open call for evidence published online by the UK House of Lords. CSER researchers facilitated collaboration and distribution of work in paper drafting using a live Google doc, and collated established and novel evidence, using diagrams to show interrelations among topics. A final White Paper was produced that was concisely presented and delivered intuitively for policy-makers.[7] When responding to follow-up queries, we assigned team members based on expertise and obtained co-authors' agreement before sending the final answers, as well as pursuing follow-up workshops with other institutions and co-creating media articles based on this work.

3.3 Drafting a parliamentary Welfare of Future Generations Bill

The All Party Parliamentary Group for Future Generations created a briefing paper as a basis for generating buy-in through tailored engagements and maintained a database of key people to be contacted (e.g. parliamentarians), including their interests/history. CSER was also able to leverage our own networks, roundtables, and events to gain key people as allies and keep in regular contact. We also worked with the campaigning and fundraising organisation Today for Tomorrow, part of The Big Issue, to support strategy for pushing the bill. Bill templates and drafting support were offered by the Parliamentarians Bill Office, and CSER researchers joined others in open drafting sessions to develop different aspects of the bill. The bill was introduced to Parliament by Lord Bird and passed through the House of Lords. However, there was not enough parliamentary time for it to pass the House of Commons as well. Nevertheless, the bill stimulated parliamentary discussions and raised the issue of the long term in UK politics.

3.4 Advising an intergovernmental organisation on foresight systems

CSER has sought out opportunities through our networks and researching the needs of organisations. We created an initial proposal by conducting a literature review on best foresight methods relevant to organisations' needs, and identified and researched relevant components of the organisation, along with the key people for conducting an expert elicitation. We also co-designed a tailored system for the World Health Organization through iterative workshops and interviews and coached the organisation through the first implementation, sharing co-authorship of two publications to get buy-in.[8]

3.5 Providing expert advice to the Cabinet Office

CSER was invited to submit proposals by leveraging our networks. We partnered with policy-bridging organisations, such as CLTR, to make connections and train us on the process, and developed recommendations backed up with substantial scientific evidence, while maintaining a database over time. We clearly defined the expectations and agenda, and established a strategy to present recommendations, even following scripts when offering recommendations to ensure that key points and messages were covered. We also paid close attention to the ensuing discussion and provided substantiated responses to follow-up questions.

3.6 Academics at the UN Biological Weapons Convention (BWC)

CSER identified shortfalls in global governance, in this case at the BWC negotiations, and defined relevant topics. We collaborated with a BWC expert on the specific process to understand what could realistically be achieved, and built trust with stakeholders through conversations to understand their expectations and perspectives. We also organised workshops with stakeholders, with each participant presenting for five minutes under Chatham House rules. We produced a report from the workshop, drafting different versions suited to particular audiences.[9] Finally, we submitted the report to UN BWC for dissemination and used it as a basis for academic/media articles.

4. Advice for Policy Engagement

Engaging in the policy process is rewarding and crucial to moving from knowledge into action and impact. It can be demanding for academics and requires a set of skills that are quite different from those developed when completing traditional academic training. When engaging with decision-shapers and decision-makers, time is of the essence. It is rare to get into academic details with most decision-makers. Communication generally needs to be concise, and when consulting one must be mindful of their competing priorities.

Monitoring the policy system is important to intervening at the right moment, when proposals are more likely to be heard and received well. When drafting reports, it is important to leave enough time when seeking feedback for drafting and redrafting, allocating time for follow-on engagement. Consulting with stakeholders before any final version of a document is a good way to involve people and give them a sense of agency within the relationship.

Achieving impact involves observing and sensing the complexities of the situation. This involves listening and understanding the needs of the stakeholder you are engaging with and responding to their concerns and perspective. Establishing a constructive relationship demands sensitivity. When seeking influence and providing expert advice, it is vital to express your message in a manner that aligns with the overall objectives and framing, that is, unless the purpose is to shift that framing (which is a very tall order). Usually, policy-makers will appreciate well-presented, concrete, actionable policy recommendations.

Collaboration amongst stakeholders with diverse interests and approaches is more likely to have an impact, even whilst such collaboration can often come with significant coordination challenges. Partners bring their own networks and constituencies, culture and messaging, credibility and perspectives. Surprising collaborators that effectively straddle polarised viewpoints can be particularly effective.

Proposals are more likely to be successful if you can demonstrate that they align with a clear consensus within the scientific community. They are also more likely to be successful if they are adaptable to diverse conditions, concerns and perspectives within the complex policy system and are expressed in the policy language and culture you are trying to influence.

It is good to communicate quantitative data, but when doing so, it is essential to be clear and unambiguous and ensure that it was understood correctly by policy-makers within the context it was situated in. If your numbers are estimates or include error bars, state this clearly. Trust is built if the uncertainties involved are communicated with clarity and assumptions exposed.

When engaging with political actors, such as members of Parliament, quality is often better than quantity, as people need to build convincing narratives if they are to use your research effectively. Spend time with individual members in one-to-one meetings, ensuring that you understand their priorities and they understand the evidence, perspective and recommendations you are conveying. Strong advocates within a body such as Parliament, who devote energy to the issue, are worth far more than a number that would simply support the idea in the lobby but not prioritise the issue.

Priorities require shifts in resources. When advocating for more resources for a case, it will have more credibility if you are able to identify other areas that could receive less.

Avoid academic jargon. Your familiarity with the use of terms and acronyms may give you comfort and a sense of expertise and solidity, but can also confuse and close down those you seek to engage. It is often helpful to have someone less familiar with your discipline to read through any outputs and check that they can clearly understand them.

5. Conclusion

Our survival in the face of existential risk demands significant shifts in activity in the public and private sectors. Policy impact takes attention, strategy, collaboration, and engagement throughout a research project. This chapter has offered targeted advice to researchers, collated from experienced CSER staff.

Notes and References

[1] The definition of policy impact according to the 2014 Research Excellence Framework is "any effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia". REF2014 was the first national assessment exercise to evaluate the wider, socioeconomic impact of research. https://www. research-strategy. admin.cam.ac.uk/files/collecting_research_impact_evidence_best_ practice_guidance.pdf

[2] NGOs, charitable organizations, schools, labour unions, indigenous groups, political parties, professional associations, foundations, faith-based organizations.

[3] Governmental departments, agencies, and organizations at local, national, regional and international levels.

[4] From startups to multinationals across a range of sectors in IT, biotechnology, finance, energy, insurance, agriculture, etc

[5] "How to Evidence and Record Policy Impact A 'how to' guide for Researchers" (2017) https://www.iph.cam.ac.uk/wp-content/uploads/2017/07/Policy-Impact-booklet-print-April-2017-1.pdf

[6] Jones, Natalie, Mark O'Brien, and Thomas Ryan. "Representation of future generations in United Kingdom policy-making", *Futures* 102 (2018): 153–63. Also reprinted as Chapter 22 of this volume.

[7] Available at https://www.cser.ac.uk/resources/written-evidence-lords-select-committee-artificial-intelligence/

[8] Available at https://www.cser.ac.uk/resources/who-emerging-technologies-and-dual-use-concerns/ and https://www.cser.ac.uk/resources/emerging-trends-and-technologies-horizon-scan-global-public-health/.

[9] Available at https://www.cser.ac.uk/resources/eighth-review-conference-biological-weapons-convention-where-next/.

19. The Cartography of Global Catastrophic Governance

Catherine Rhodes and Luke Kemp

Highlights:

- This chapter provides an overview of the fragmented and insufficient international governance arrangement for GCR hazards and drivers.
- It finds that despite clusters of dedicated regulation and action — including in nuclear, chemical, and biological warfare, climate change, and pandemics — their effectiveness is often questionable.
- In other areas, such as catastrophic uses of AI, asteroid impacts, solar geoengineering, unknown risks, super-volcanic eruptions, inequality and many areas of ecological collapse, the legal landscape is littered more with gaps than effective policy.
- The authors suggest five steps to help advance the state of GCR governance: 1) identifying instruments and policies that can address multiple risks and drivers; 2) researching the relationship between drivers and hazards to create a deeper understanding of "civilisational boundaries"; 3) exploring the potential for "tail risk treaties" that swiftly ramp-up action in the face of early warning signals of catastrophic change; 4) examining the coordination and conflict between different GCR governance areas; and 5) building the foresight and coordination capacities of the UN for GCR.

https://doi.org/10.11647/OBP.0360.19

- These recommendations can ensure that international governance navigates the turbulent waters of the 21st century, without blindly sailing into the storm.

This chapter was written by CSER researchers for the Global Challenges Foundation and provides an overview of existing governance frameworks. For a proposal on how to improve the global governance of nuclear weapons, see Chapter 20. For a methodological framework that can help identify the early warning signals required to make anticipatory governance like tail risk treaties work, see Chapter 17.

1. Introduction

On January 24th 2019, the fingers on the Doomsday Clock did not move: they stayed pressed ominously at two minutes to midnight. The clock has been the most captivating attempt to forecast the likelihood of a Global Catastrophic Risk (GCR). It is inherently limited, focusing only on a subset of GCRs: nuclear weapons, climate change and more recently epistemic security. It also does not reflect the governance of different global risks. Understanding how humanity is currently responding to GCRs is fundamental in comprehending how precarious or resilient the world is to calamity.

While Global Catastrophic Risks are becoming increasingly widely known, their governance is understudied. Only a handful of studies have examined whether existing international law arrangements,[1] or the UN,[2] are fit for addressing existential or Global Catastrophic Risks. Others have attempted to look at the capability of the UN to prevent new risks in an age of AI and converging, powerful technologies.[3] These studies have relied on more of a cursory overview of governance, focusing on broad structures and scenario analysis. They have not systematically examined coverage of different hazards and vulnerabilities.

Our report seeks to overcome these limitations by providing the most far-reaching and comprehensive mapping of the governance of Global Catastrophic Risks, including both hazards and vulnerabilities. Our definition of GCRs and existential risks is provided below in Table 1. While our report will focus on GCRs broadly, many of the assessed issues are plausible of becoming existential risks as well.

Table 1: Definitions of GCRs and existential risks.

Term	Definition
Existential risk	Any risk that has plausible pathways to cause either human extinction or the drastic and permanent curtailment of societal progress.[4] A global collapse could be considered as a lower bound for this, given the uncertainty of how it would unfold in the presence of weapons of mass destruction.[5]
Global Catastrophic Risk	Any risk that plausibly leads to the loss of 10% or more of global population.[6]

Our *Cartography of GCRs* demonstrates that several GCR hazards (climate change, nuclear weapons) are covered by international law but usually inadequately. That is, the institutions often lack clear enforcement and compliance mechanisms, and have largely failed to address the underlying collective action problem. Other issues, such as solar geoengineering, catastrophic uses of AI, inequality and some areas of ecological collapse (phosphorous, nitrogen and atmospheric aerosols), are either largely or completely neglected. The governance across GCRs is fragmented, with fractured membership and mandates both within and across different hazards and vulnerabilities. There is no central body empowered to coordinate responses to GCRs nor to foresee them.

2. Approach

In order to achieve a comprehensive overview of global governance arrangements for GCRs it is important to adopt a broad conception of global governance, because otherwise key components may be overlooked, and indications of emergent activity may not be apparent.

A core focus of research and practice in global governance — which is also reflected in this report — justifiably remains the actions of states through international (intergovernmental) organisations and international legal instruments. A report that only focused on these components would, however, present an incomplete picture: a range of other intergovernmental governance activities can contribute to addressing GCRs; and there are relevant activities outside the intergovernmental space. For some GCR areas, the latter currently dominate global governance arrangements.

The significance of different components varies between GCR regimes. This means that the construction of maps and attention paid to different components varies too, but we have also aimed for a level of consistency in presentation. For example: we cover bilateral agreements more extensively in nuclear warfare than in other areas because of their high significance in managing global nuclear risk; we cover multilateral expert communities extensively in the Asteroid Impact and Super-Volcanic Eruption areas, because these are more heavily relied upon there.

It is worth making a general observation about the increasing range of issues that need to be addressed through global governance and the challenges this presents:

- Formal intergovernmental governance activities are generally poorly resourced already; their capacity to take on additional tasks and remain responsive to new threats is limited, and some are already overstretched.
- Proliferation of global governance activities can disadvantage less well-resourced states, which can struggle to participate in a large number of international forums and processes, representativeness in which is already sub-optimal.
- Increased complexity generally makes governance arrangements more difficult to navigate (one of the reasons mapping work is useful) and increases the transaction costs associated with international cooperation, the likelihood of conflicts and contradictions between rules, and duplication of effort.

Given the extent and complexity of many of the regimes covered in this report, we have separated some more detailed information into Appendix I. Appendix II provides a list of acronyms. These appendices are available online.

(https://globalchallenges.org/app/uploads/2023/06/The-Cartography-of-Global-Catastrophic-Governance-2019.pdf)

In the report itself, we provide maps and summary information for individual hazards in global GCR governance.

These generally follow the GCR categories from GCF's *Global Catastrophic Risks 2018 Report*. The areas of "Biological and Chemical Warfare" and "Pandemics" have been combined, because there are significant overlaps in the global governance activities across these areas that are best illustrated by handling them together. We have also designated "Ecological Collapse" as a driver of GCR, rather than a hazard. Otherwise we have consistently applied the categorisation of the 2018 report.[7]

The mapping of each of these areas is intended to be representative but not exhaustive. We instead provide an overview of key treaties and governance efforts and characterise these as a regime complex: a constellation of institutions addressing the same international issue.[8] We provide information about the gaps and issues requiring attention in each regime at the end of each hazard section. We also deliver a high-level view of broader GRC governance arrangements under the UN and transnational (networks of non-state actors) actions.

We end with a summary assessment of GCR governance arrangements and identification of (priority) lines of research and practical action that could advance the governance of individual GCRs and GCRs collectively.

3. Regime Complexes for Hazards

Hazards are direct threats that could cause global calamity. We draw on both previous GCR reports, as well as consultations with our colleagues to produce the following list of relevant hazards: AI; Asteroid Impact; Pandemics, Biological and Chemical Warfare; Climate Change; Solar Geoengineering; Unknown Risks; Nuclear Warfare, and Super-Volcanic

Eruptions. We provide a high-level summary of the governance arrangements of each of these hazards before concluding with an analysis of their effectiveness and gaps.

3.1 AI

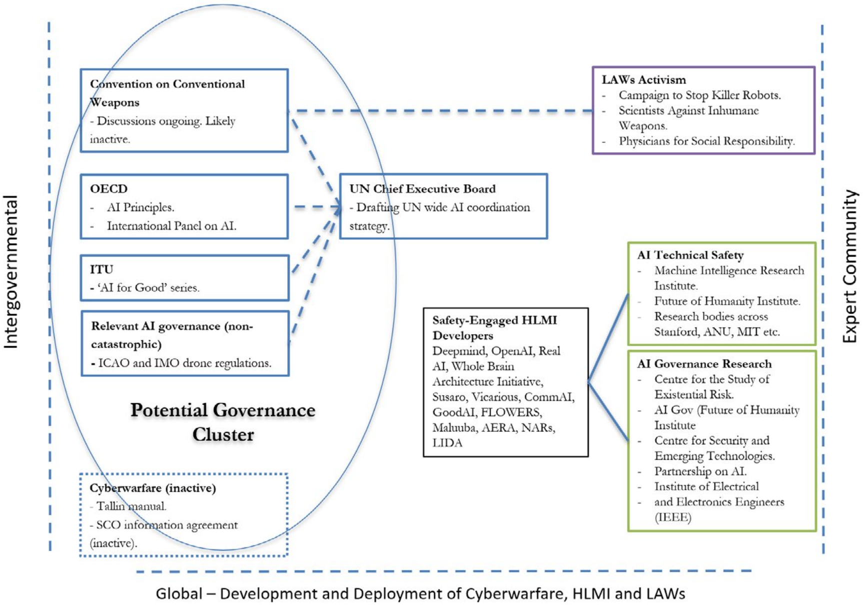

Intergovernmental
Expert Community
Convention on Conventional Weapons
- Discussions ongoing. Likely inactive.
OECD
- AI Principles.
- International Panel on AI.
ITU
- 'AI for Good' series.
Relevant AI governance (non-catastrophic)
- ICAO and IMO drone regulations.
Potential Governance Cluster
Cyberwarfare (inactive)
- Tallin manual.
- SCO information agreement (inactive).
UN Chief Executive Board
- Drafting UN wide AI coordination strategy.
LAWs Activism
- Campaign to Stop Killer Robots.
- Scientists Against Inhumane Weapons.
- Physicians for Social Responsibility.
Safety-Engaged HLMI Developers
Deepmind, OpenAI, Real AI, Whole Brain Architecture Initiative, Susaro, Vicarious, CommAI, GoodAI, FLOWERS, Maluuba, AERA, NARs, LIDA
AI Technical Safety
- Machine Intelligence Research Institute.
- Future of Humanity Institute.
- Research bodies across Stanford, ANU, MIT etc.
AI Governance Research
- Centre for the Study of Existential Risk.
- AI Gov (Future of Humanity Institute
- Centre for Security and Emerging Technologies.
- Partnership on AI.
- Institute of Electrical
- and Electronics Engineers (IEEE)
Global – Development and Deployment of Cyberwarfare, HLMI and LAWs

Fig. 1: Catastrophic AI regime complex.

Within the rising age of AI are hidden disastrous developments. There is an open debate over whether AI systems as a class can be regulated. This is because AI is a set of techniques and sub-disciplines rather than a single, specific technology.[9] However, there are certain, specific forms of AI systems and end uses which could constitute a GCR. These form a discernible, governable cluster. These include:

- AI-enabled cyberwarfare;
- The creation of a misaligned or misused "High Level Machine Intelligence" (HLMI): a generalised AI system that is roughly equivalent to a human in its cognitive capabilities;
- Lethal autonomous weapons.

Cyberwarfare has essentially no governance at the international stage. There are two minor exceptions. First, is the Tallinn Manual on the International Law Applicable to Cyber Warfare. The Tallinn Manual has only been endorsed by NATO member states and provides non-binding advice on the application of international law to cyberspace. Second, is the Shanghai Cooperation Organisation's "Information Security Agreement". This has only six member states and failed to garner sufficient approval from the UN General Assembly. The absence of effective regulation and the proliferation of threats has led some to call for a Cyberwar Convention.[10] Negotiations for such a body have not begun and are not on the horizon.

LAWs could potentially be covered under the Convention on Certain Conventional Weapons. The Convention has a mechanism — its Additional Protocols — to expand its coverage to new categories of weapons (such as blinding lasers or land mines). However, in practice, negotiations to include LAWs under its remit have been marked by disinterest from great powers. It has yet to yield any success and appears unlikely to do so in the foreseeable future. If it did, the Convention has no ability to enforce its decisions. In the absence of effective international law, civil society has stepped forward in the form of the active Campaign to Stop Killer Robots.

The development of HLMI is ungoverned. It is the most neglected area of international AI law.[11] In the absence of explicit regulation, both corporate self-governance and expert community action have filled the void. Many of the firms and bodies creating HLMI are actively engaged in safety work. One 2017 survey of 45 HLMI projects across 30 countries and six continents found that only 15 were directly involved in AI safety research.[12] Many of these are directly connected to academic and civil society groups working directly on AI technical safety or AI governance. Bodies such as CSER and AI Gov (under the Future of Humanity Institute at Oxford University) are all actively engaged with prominent HLMI developers such as Deep Mind (part of Google) and OpenAI.

There is also no consensus on the governing principles for AI systems. The work of both these expert communities and others has spawned a plethora of AI principles. Most of these encapsulate some common, ambiguous concepts: use of AI for the common good; avoiding harm and the infringement of rights; and privacy, fairness and autonomy. No clear set of principles reigns supreme, and several tensions exist across them.[13] It is unclear how directly or effectively any of these is for catastrophic AI applications specifically.

There are also several bodies that have some relevance to AI systems but no direct mandate over them. The ITU has been admirably active in promoting AI dialogue through hosting annual "AI for Global Good Summits" since 2017. Yet the ITU is currently limited to regulating telecommunication systems, such as radio infrastructure; efforts to expand its role in internet governance have been resisted. There are legal arguments that its mandate could extend over many AI systems, but this seems politically unlikely to happen. Similarly, the International Organisation for Standardization (ISO) has established a committee to discuss a programme on AI standards, but would have no mandate to address the identified AI problems on its own.

Alongside these bodies is a raft of regulations, working groups and decisions under other fora. Action across the IMO, ICAO, ITU, and other bodies, as well as treaty amendments, such as the updating of the Vienna Convention on Road Traffic to encompass autonomous vehicles, are indicative of this.[14] Most recently, France and Canada have jointly led an initiative to establish a 'International Panel on AI' under the OECD. This proliferating panoply of AI governance shows some signs of self-organising. The UN System Chief Executives Board (CEB) for Coordination through the High-Level Committee on Programmes has been empowered to draft a system-wide AI engagement strategy. Whether such coordination will be successful is unclear. Moreover, this swell of governance does not capture the catastrophic uses cyberwar, LAWs and HLMI.

Table 2: Coverage and gaps in HLMI governance.

Coverage	Expert communities and civil society have been increasingly active in campaigns against LAWs, as well as technical and governance research on HLMI.
Gaps	HLMI currently has no direct governance under international law. LAWs falls under the mandate of Convention on Conventional Weapons but has not been regulated to date. Similarly, attempts to govern cyberwarfare have been either plurilateral and non-binding (Tallin Manual) or unsuccessful (SCO Information Agreement).

Issues requiring attention	Whether and how these issues could be addressed in-tandem, such as through a body focused on the military applications of AI. The legitimacy and potential dangers of self-regulation focused HLMI development.

3.2 Asteroid impact

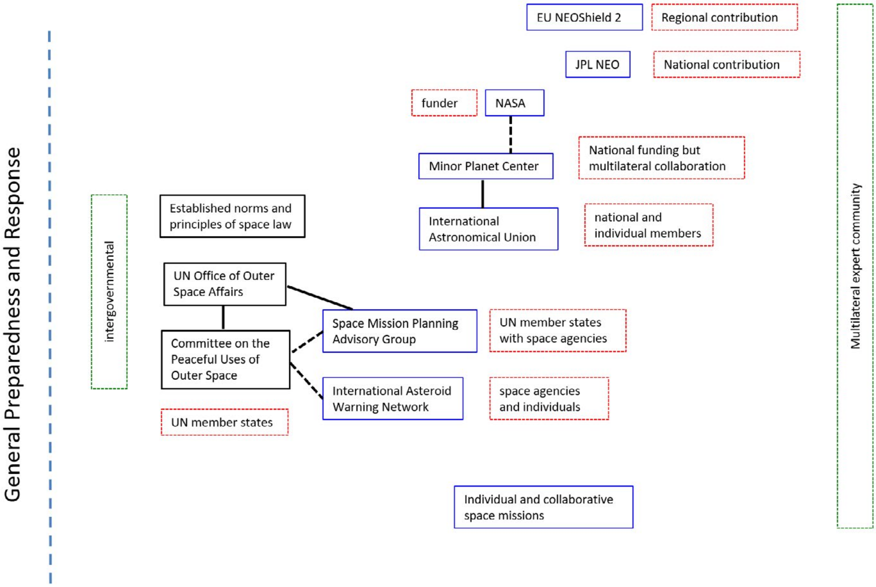

EU NEOShield 2
Regional contribution
JPL NEO
National contribution
funder
NASA
Minor Planet Center
National funding but multilateral collaboration
International Astronomical Union
national and individual members
Established norms and principles of space law
UN Office of Outer Space Affairs
Committee on the Peaceful Uses of Outer Space
Space Mission Planning Advisory Group
UN member states with space agencies
International Asteroid Warning Network
space agencies and individuals
UN member states
Individual and collaborative space missions
intergovernmental
Multilateral expert community
General Preparedness and Response

Fig. 2: Asteroid impact regime complex.

Compared to most other GCRs, global governance for asteroid impacts is minimal, and not particularly complex. There is a reasonable quality of coverage for the more technical aspects of identification, monitoring, evaluation, and early warning, as well as coordination and promotion of research, development and testing of deflection techniques. (Broadly, all of those activities focus on prevention.) There is some coordination of planning around communication, for scenarios in which a "credible impact threat" is identified, and some connections with civil defence communities (for example, as part of the response activities of the Space Mission Planning Advisory Group).

Participants in these governance arrangements understand the seriousness of the threat, particularly where an NEO would be large enough to directly cause a global cooling effect (>1km), and an understanding that smaller NEOs (in the 140m-1km range) could indirectly have global catastrophic impacts as well as being locally catastrophic. There is clear hope that there will be sufficient warning time in advance of a significant Earth impact to boost resilience efforts; however, there is limited extension of the NEO-specific governance arrangements to address preparedness and response. Mostly this will depend on more general global governance arrangements for disaster preparedness and response (see Section 5). Notably, the severe impacts that would need to be prepared for and responded to — those associated with the effects of global cooling and damage to critical infrastructure) will be very similar to those caused by some other GCRs, such as super-volcanic eruptions (Section 3.8) and nuclear winter scenarios (Section 3.7).

While the UN Committee on Peaceful Uses of Outer Space (COPUOS) and Office on Outer Space Affairs (UNOOSA) are at the core of global governance of asteroid impacts, most of the governance efforts are undertaken by scientific and technical experts in national space agencies, research institutions, and through individual contributions. Some national (particularly NASA-funded) and regional (e.g. the EU's NEOShield 2 Project) efforts have particular significance.

The activities of these other groups connect back to COPUOS and strongly emphasise openness, sharing of data and analysis, and collaborative efforts. This arrangement seems to function well for addressing the technical and prevention aspects of asteroid impact governance; however, attention is needed for sustainability and continuity should, for example, a major partner withdraw. (Ensuring continuity has, for example, motivated the establishment of the UNOOSA as a permanent secretariat for the Space Mission Planning Advisory Group.)

Issues around representativeness and equity might in future arise in this governance area, but — currently at least — this seems much less problematic than in other GCR governance areas (such as pandemics), particularly when focusing on the technical and preventative aspects. For representativeness, while the Space Mission Planning Advisory Group, for example, requires the ability to contribute to space missions for participation, and is therefore oriented towards states with space agencies, COPUOS is open to all UN member states (92 are currently members of the Committee) and its recommendations go to the UN General Assembly for discussion and approval. Thus, all UN member states have an opportunity to engage with its work.

For equity, core principles of space law — benefit to humanity and non-appropriation — are established across this governance regime and appear to have broad acceptance and strong normative force. COPUOS has programmes relating to capacity building in space law and for application of space technologies for development goals and during disasters.

It is expected that technological advances will enable mining of NEOs for resources at some point in the future, most likely for use in outer space rather than return to Earth. If this area is substantially financed and/or operated by commercial enterprises, then the practicalities of benefit-sharing will need further consideration. COPUOS will be an appropriate forum for such discussion. COPUOS is also an appropriate point for connection with institutions in the general disaster preparedness and response areas of global GCR governance. The International Asteroid Warning Network (IAWN) is currently working on definitions and terminology for NEOs, and this will include definition of NEO as a natural hazard to feed into the UN Office on Disaster Risk Reduction's updated glossary of natural hazards.[15]

Table 3: Coverage and gaps in asteroid impact prevention and preparedness.

Coverage	There is a good level of coverage for: identification, observation, monitoring, analysis and evaluation, communication, and preventative response. It is limited for: impact preparedness, resilience and response — quality of coverage of these areas will therefore largely depend on general disaster preparedness and response efforts.
Gaps	These are likely to be found in the general disaster preparedness and response efforts.
Issues Requiring Attention	Sustainability and continuity (particularly of non-intergovernmental arrangements). Increasing representativeness and engagement. Increasing role of commercial enterprises.

3.3 Pandemics, biological and chemical warfare

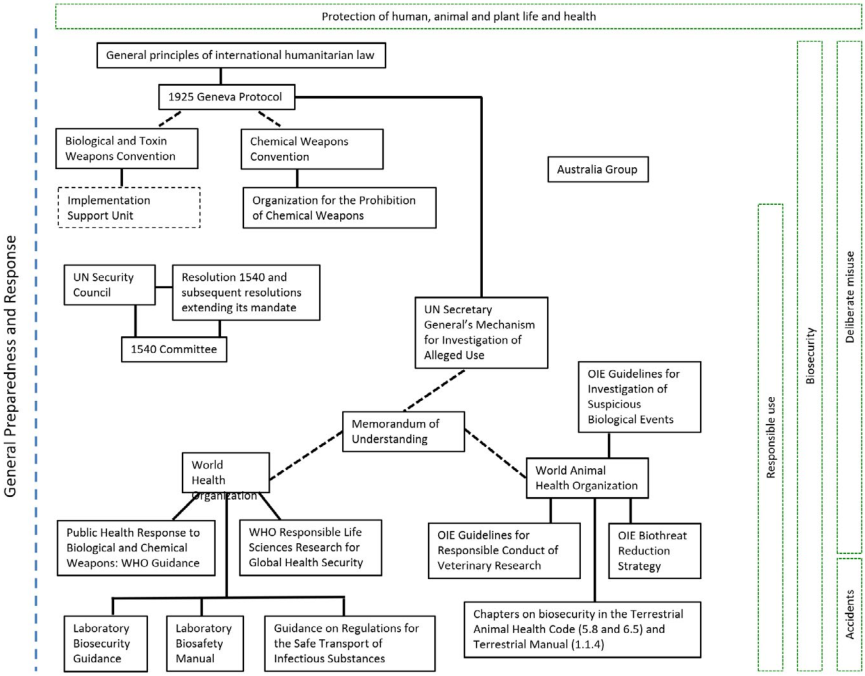

Protection of human, animal and plant life and health
General Preparedness and Response
General principles of international humanitarian law
1925 Geneva Protocol
Biological and Toxin Weapons Convention
Chemical Weapons Convention
Implementation Support Unit
Organization for the Prohibition of Chemical Weapons
Australia Group
UN Security Council
Resolution 1540 and subsequent resolutions extending its mandate
1540 Committee
UN Secretary General's Mechanism for Investigation of Alleged Use
OIE Guidelines for Investigation of Suspicious Biological Events
Memorandum of Understanding
World Health Organization
World Animal Health Organization
Public Health Response to Biological and Chemical Weapons: WHO Guidance
WHO Responsible Life Sciences Research for Global Health Security
OIE Guidelines for Responsible Conduct of Veterinary Research
OIE Biothreat Reduction Strategy
Laboratory Biosecurity Guidance
Laboratory Biosafety Manual
Guidance on Regulations for the Safe Transport of Infectious Substances
Chapters on biosecurity in the Terrestrial Animal Health Code (5.8 and 6.5) and Terrestrial Manual (1.1.4)
Responsible use
Biosecurity
Deliberate misuse
Accidents

Fig. 3: Pandemics, biological and chemical warfare regime Complex I.

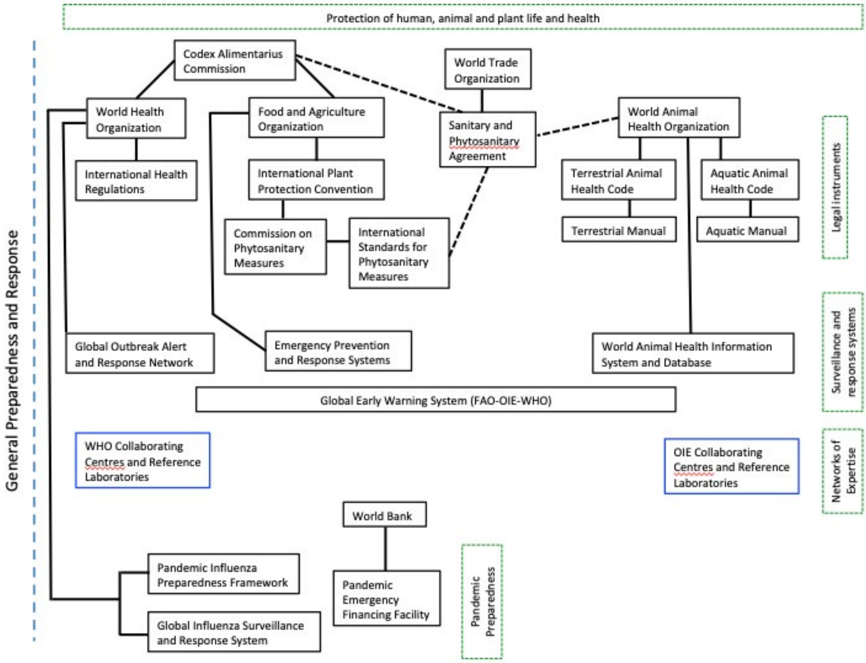

Protection of human, animal and plant life and health
General Preparedness and Response
Codex Alimentarius Commission
World Trade Organization
World Health Organization
Food and Agriculture Organization
Sanitary and Phytosanitary Agreement
World Animal Health Organization
International Health Regulations
International Plant Protection Convention
Terrestrial Animal Health Code
Aquatic Animal Health Code
Commission on Phytosanitary Measures
International Standards for Phytosanitary Measures
Terrestrial Manual
Aquatic Manual
Legal instruments
Global Outbreak Alert and Response Network
Emergency Prevention and Response Systems
World Animal Health Information System and Database
Surveillance and response systems
Global Early Warning System (FAO-OIE-WHO)
WHO Collaborating Centres and Reference Laboratories
OIE Collaborating Centres and Reference Laboratories
Networks of Expertise
World Bank
Pandemic Influenza Preparedness Framework
Pandemic Emergency Financing Facility
Global Influenza Surveillance and Response System
Pandemic Preparedness

Fig. 4: Pandemics, biological and chemical warfare regime Complex II.

In this summary we combine consideration of global governance of biological and chemical warfare and pandemics, because there are significant areas of overlap between the governance arrangements for these two areas, which might not be fully apparent when addressing them separately.

The range of biological risks addressed by global governance is illustrated by the World Health Organization’s “biorisk spectrum”:

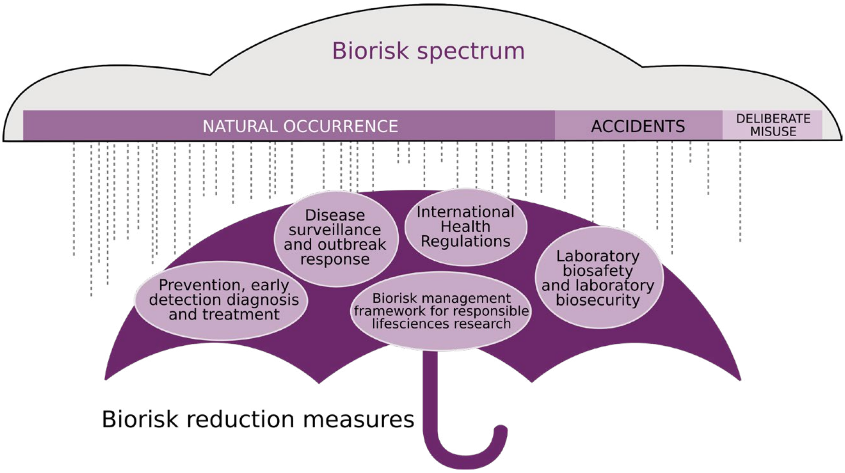
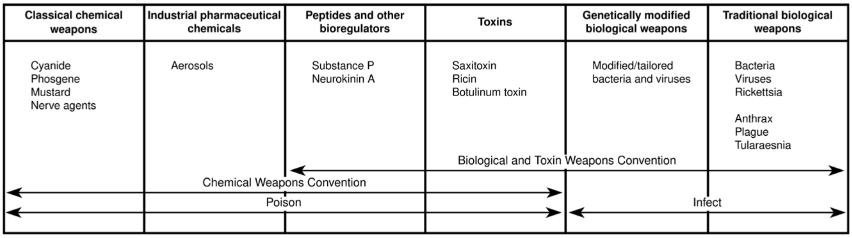

Biorisk spectrum
NATURAL OCCURRENCE
ACCIDENTS
DELIBERATE MISUSE
Disease surveillance and outbreak response
International Health Regulations
Laboratory biosafety and laboratory biosecurity
Prevention, early detection diagnosis and treatment
Biorisk management framework for responsible lifesciences research
Biorisk reduction measures

Fig. 5: The biorisk spectrum and biorisk reduction measures.[16]

To this, it is worth adding two further categories to the spectrum: "human-induced" lies between natural occurrence and accidents, and would for example cover anti-microbial resistance as a threat that is "natural" but driven primarily by human action, and might also cover e.g. shifts in geographical range of disease vectors driven by climate change; and "deliberate action with benign intent but unintended consequences" which would sit between accidents and deliberate misuse. This might, for example, relate to release of a biological control agent into the environment without understanding its consequences for health. While this particular image focuses on human health (as the responsibility of the WHO), there are Global Catastrophic Biological Risks associated with threats to animal and plant health, and to ecosystems — particularly where these would severely impact food safety and security and key ecosystem services.

Another risk spectrum to be aware of is that which extends across biological and chemical warfare:



Classical chemical weapons	Industrial pharmaceutical chemicals	Peptides and other bioregulators	Toxins	Genetically modified biological weapons	Traditional biological weapons
Cyanide Phosgene Mustard Nerve agents	Aerosols	Substance P Neurokinin A	Saxitoxin Ricin Botulinum toxin	Modified/tailored bacteria and viruses	Bacteria Viruses Rickettsia Anthrax Plague Tularaesnia

Biological and Toxin Weapons Convention
Chemical Weapons Convention
Poison
Infect

Fig. 6: The comprehensive prohibition of the chemical weapons convention and the biological and toxin weapons convention.[17]

This illustrates the areas of overlapping coverage between the two conventions. While there are now separate conventions for biological and chemical weapons, they were initially addressed together in international governance, and there remain significant connections between the two regimes. The 1925 Geneva Protocol prohibits use of biological and chemical agents in war. It still has relevance because the prohibition on development, production, stockpiling, acquisition and retention in the Biological and Toxin Weapons Convention (BTWC) extends to use through reference to the Geneva Protocol, and because the Protocol is accepted as part of customary international law applicable to all states whether or not they are party to the conventions.

The BTWC and Chemical Weapons Convention (CWC) utilise general purpose criteria prohibiting use of biology and chemistry for non-peaceful purposes. States parties to the conventions have repeatedly emphasised that they are applicable to all scientific and technological advances in relevant fields. Both conventions include provisions promoting peaceful applications — for the BTWC "prevention of disease" is specifically mentioned in this regard, and this is one way in which they connect with other areas of governance of biological risks.

The long-standing international norms against biological and chemical weapons have experienced some challenges, but while there is some concern around potential erosion, these remain strong at present and are central to global governance efforts. There are also some well-recognised areas of weakness in the conventions. The CWC's provisions relating to the permitted use of some toxic chemicals for law enforcement purposes, has resulted in some ambiguities and divergent interpretations — for example, about development and use of riot control agents and incapacitants.[18] The CWC is overseen by the Organization for the Prohibition of Chemical Weapons, which has around 500 staff and an annual budget of around €70 million. One of its core roles is verification activities, which are structured around inspection regimes. The BTWC does not have an associated international organisation, and is instead supported by a small Implementation Support Unit of three staff. It also has no verification regime (attempts to negotiate one failed in the early 2000s and are yet to be re-established). This is a significant weakness given the dual-use nature of biological facilities, equipment, materials and research. Both conventions cover areas of rapid scientific and technological advance and their effective implementation by states parties needs to be informed by a good understanding of the risks and opportunities associated with such advances. The OPCW has a Science Advisory Board that undertakes some of this work in regard to the CWC. This is, however, another area in which the BTWC has extremely limited capacity. Civil society groups such as research institutions and science academies undertake efforts in support of science and technology review for the conventions. These efforts are important, but can lack some of the legitimacy of formal processes.

Other international governance relevant to deliberate misuse includes: UN Security Council Resolution 1540(2004), which addresses potential proliferation of biological, chemical and nuclear weapons to non-state actors, and subsequent resolutions which extended its mandate,[19] and the associated 1540 Committee, which reports to the Security Council on its implementation; and the UN Secretary General's Mechanism for Investigation of Alleged Use of Chemical and Biological Weapons.

The OPCW and ISU undertake some activities to support assistance in case of a biological or chemical weapons attack, including through facilitation of requests and offers by their states' parties. OPCW has also produced a *Practical Guide for Medical Management of Chemical Warfare Casualties*, directed to medical responders, and the WHO also provides relevant advice, including in its *Public Health Response to Biological and Chemical Weapons* guidance.

There are two other key overlapping areas with broader global governance of biological risks. First, measures for laboratory biosafety and biosecurity, and safety during transport of infectious materials, which form part of the work of the World Health Organization (WHO) and World Animal Health Organization (OIE) contribute to the safeguarding of biological materials that might be misused. Secondly, the systems for surveillance, preparedness and response to disease events overseen by the WHO, OIE and Food and Agriculture Organization (FAO) will play a key role in detection and response to any deliberate disease outbreaks or chemical attacks. OIE and WHO both have memorandums of understanding around provision of technical support with the UN Secretary-General's Mechanism for Investigation of Alleged Use.

FAO, OIE and WHO also play important roles in prevention and response to accidental releases of biological agents, toxins and hazardous chemicals, including specific guidance on safety in laboratories and during transport.[20] Their general surveillance, preparedness and response systems will play a key role in detection and response to any outbreaks resulting from accidents or deliberate releases with benign intent but unintended consequences. Provisions of the Convention on Biodiversity and its Cartagena Protocol on Biosafety may also have relevance where damage to health or the environment stems from transboundary movements of living modified organisms.

WHO and OIE have also produced some guidance (*Responsible Life Sciences Research for Global Health Security*; and *Guidelines for Responsible Conduct of Veterinary Research: Identifying, Assessing and Managing Dual-Use*) that is complementary to BTWC states parties' discussions and decisions promoting education and training of scientists in biosecurity responsibilities.

The main international organisations responsible for protection of human, animal and plant life, and health (and therefore for addressing threats to them) are the WHO, OIE and FAO. The WHO and FAO also jointly established the Codex Alimentarius Commission to work on international food and feed safety. The disease control activities of each organisation centre around specific legal instruments:

- The International Health Regulations (2005);
- The Terrestrial Animal Health Code, Manual of Diagnostic Tests and Vaccines for Terrestrial Animals, Aquatic Animal Health Code, and Manual of Diagnostic Tests for Aquatic Animals; and

- The International Plant Protection Convention.

Their work is also supported by surveillance and response systems, expert advisory groups and networks, and collaborating centres and laboratories. The WHO, for example, has over 800 collaborating centres in 80 countries supporting its programmes, and the OIE has 60 collaborating centres, and a network of reference laboratories focusing on scientific and technical research on over 100 serious animal diseases. Surveillance and response activities, include generalised systems such as the Global Outbreak Alert and Response Network, World Animal Health Information System, and FAO's emergency prevention and response systems (EMPRES); and disease specific systems such as the WHO's Global Influenza Surveillance and Response System.

In response to a breakdown in the international system for sharing of influenza viral samples in 2006/2007, the WHO took action to revise its Global Influenza Surveillance Network, enhancing traceability through an Influenza Virus Traceability Mechanism, and establishing the Pandemic Influenza Preparedness Framework, which includes centralised stockpiles of vaccines and treatments for distribution to developing countries during outbreaks of human pandemic potential. The Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of the Benefits Arising From Their Utilization (to the Convention on Biological Diversity) also has relevance to the international sharing of microbial genetic resources, which may interact with global public health efforts.[21]

In recognition of the overlaps between protection of human, animal and plant life and health, the FAO, OIE and WHO have instituted several cooperative initiatives, including (for example): OFFLU a FAO-OIE network of expertise on animal influenzas, and the FAO-OIE-WHO Global Early Warning System for Health Threats and Emerging Risks at the Human-Animal-Ecosystems Interface (GLEWS). They also regularly send representatives and provide information to BTWC meetings.

In general, capacity building efforts that focus on building national health system capacities will increase the effectiveness of surveillance and response efforts and reduce the risk of international spread of serious disease outbreaks. Such efforts are supported by states parties to the BTWC, WHO, OIE, FAO among other international organisations and through mechanisms such as the Standards and Trade Development Facility — a partnership between FAO, OIE, WHO, the World Bank and

World Trade Organization — that supports access to international markets through development capacities to meet and maintain international standards in food safety, animal and plant health. The World Bank has also increased its activities relating to pandemics over the last few years, including creating a Pandemic Emergency Financing Facility to support countries' outbreak response and limit their international spread.

While these activities appear extensive, there are particular concerns about their effectiveness in relation to capacity to contain and address serious outbreaks of international concern, whatever their origin. The Global Health Security Index — a partnership of the Nuclear Threat Initiative, John Hopkins Center for Global Health Security, and Economist Intelligence Unit — which focuses on assessing global health security capacities, has recently reported and raised the following key points in this regard:[22]

1. National health security is fundamentally weak around the world. No country is fully prepared for epidemics or pandemics, and every country has important gaps to address.
2. Countries are not prepared for a globally catastrophic biological event.
3. There is little evidence that most countries have tested important health security capacities or shown that they would be functional in a crisis.
4. Most countries have not allocated funding from national budgets to fill identified preparedness gaps.
5. More than half of countries face major political and security risks that could undermine national capability to counter biological threats.
6. Most countries lack foundational health systems capacities vital for epidemic and pandemic response.
7. Coordination and training are inadequate among veterinary, wildlife, and public health professionals and policymakers.
8. Improving country compliance with international health and security norms is essential.

Table 4: Coverage and gaps in pandemic preparedness and biological security governance.

Coverage	The breadth of coverage in this area is good: extending across harms to human, animal and plant health and the environment arising from deliberate misuse, accidental release, and natural occurrence of disease. The points of intersection between these areas are also reasonably well covered, and cooperative activity in those areas is increasing. However, there are some significant weaknesses within individual areas and gaps in capacity. There is a good level of engagement of expert communities in the overall work of the OIE and WHO.
Gaps	Significant gaps include: lack of verification for the BTWC; limited capacity for science and technology review for the BTWC; and in national capacities to respond to and contain outbreaks with potential for global spread (such as the core capacities required by the WHO's International Health Regulations). Pandemic preparedness capabilities in particular have been assessed as inadequate by several organisations. This is compounded by the tendency for states to prioritise protection of their own populations above effective global responses (as demonstrated during the the 2009 H1N1 influenza outbreak).
Issues requiring attention	Particular priority issues include: the need to enhance the ability of international institutions to form good understanding of emerging threats (and opportunities) associated with rapid advances in science and technology, and to adapt governance arrangements to respond effectively to them; and the need for effective action to build global capacities to respond to human pandemic threats and serious disease threats to animals and plants.

3.4 Climate change

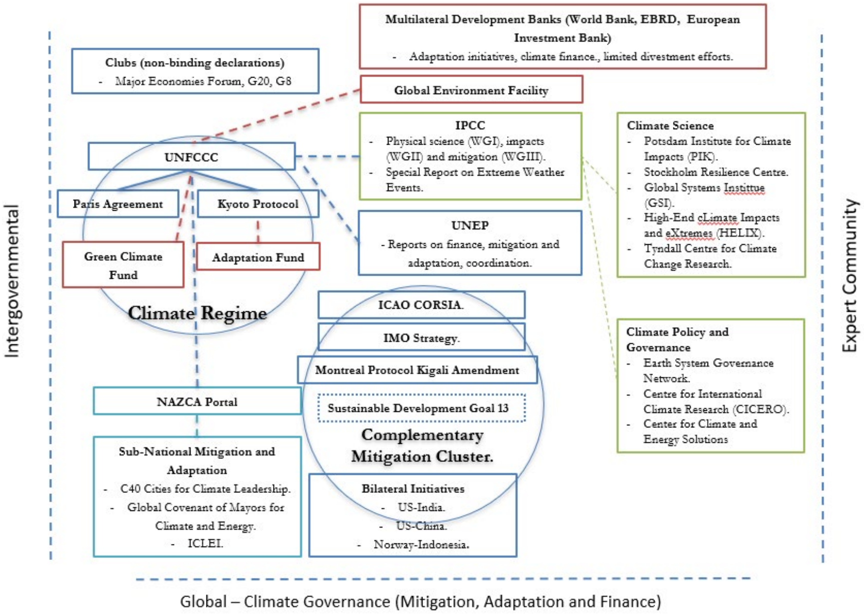

Multilateral Development Banks (World Bank, EBRD, European Investment Bank)
- Adaptation initiatives, climate finance., limited divestment efforts.
Clubs (non-binding declarations)
- Major Economies Forum, G20, G8
Global Environment Facility
IPCC
- Physical science (WGI), impacts (WGII) and mitigation (WGIII).
- Special Report on Extreme Weather Events.
Climate Science
- Potsdam Institute for Climate Impacts (PIK).
- Stockholm Resilience Centre.
- Global Systems Institue (GSI).
- High-End cLimate Impacts and eXtremes (HELIX).
- Tyndall Centre for Climate Change Research.
UNFCCC
Paris Agreement
Kyoto Protocol
Green Climate Fund
Adaptation Fund
Climate Regime
UNEP
- Reports on finance, mitigation and adaptation, coordination.
ICAO CORSIA.
IMO Strategy.
Montreal Protocol Kigali Amendment
Sustainable Development Goal 13
Complementary Mitigation Cluster.
Climate Policy and Governance
- Earth System Governance Network.
- Centre for International Climate Research (CICERO).
- Center for Climate and Energy Solutions
NAZCA Portal
Sub-National Mitigation and Adaptation
- C40 Cities for Climate Leadership.
- Global Covenant of Mayors for Climate and Energy.
- ICLEI.
Bilateral Initiatives
- US-India.
- US-China.
- Norway-Indonesia.
Intergovernmental
Expert Community
Global – Climate Governance (Mitigation, Adaptation and Finance)

Fig. 7: The climate regime complex.

The global governance of climate change is one of the most well-studied and addressed GCRs under international law. International efforts to address to climate change largely began in 1992 with the creation of the United Nations Framework Convention on Climate Change (UNFCCC). The UNFCCC has since been the lynchpin of international legal efforts to address climate change. It includes provisions on adaptation to climate impacts, mitigation, as well as broader considerations such as capacity building. It also establishes the overarching norms and principles of climate diplomacy, such as "common but differentiated responsibilities".

The UNFCCC is the focal point of the climate regime and has been operationalised through two separate protocols:

- *The Kyoto Protocol*: created in 1997, before entering into force in 2005. The Kyoto Protocol contains provisions for monitoring, transparency and verification of emissions, market-based mechanisms (including for international emissions trading and offsetting), financing, and adaptation actions and mitigation targets. It is composed of a two-annex system whereby developing country parties are bound to legally binding emissions reductions targets. Developing countries are not bound by any mitigation targets. The first commitment period of the protocol lasted until 2013. The 2012 Doha Amendment which extends to the Kyoto Protocol's second commitment period through to 2020 has yet to enter into force due to a lack of ratifying countries.

- *The Paris Agreement*: created in 2015, entered into force in October 2016. The agreement contains provisions on adaptation, mitigation, market-based mechanisms, loss and damages from climate impacts and multiple other mechanisms. The agreement has set an international target to limit global warming to well below 2 °C above pre-industrial levels and pursue efforts to keep it to 1.5 °C. It is a pledge-and-review agreement in which countries offer self-determined pledges (nationally determined contributions/NDCs) which are collectively reviewed every five years.[23] The agreement only offers one additional binding legal obligation to the UNFCCC: to put forward a pledge every five years. Its structure was watered down to allow for the US to join via an executive agreement rather than Senate ratification.[24]

These three institutions constitute the UN climate regime. They have set the primary targets and rules for adaptation and mitigation that other institutions follow and implement. In addition to adaptation and mitigation, there is also governance of loss and damages. This refers to managing the damages incurred by the detrimental impacts of climate change, including slow-onset events, and extreme weather events. In 2013 the Warsaw International Mechanism for Loss and Damage associated with Climate Change Impacts (Loss and Damage Mechanism) was established to govern this area. It offers a dialogue platform for relevant stakeholders and aims to enhance knowledge of risk management and support through finance, technology and capacity building. It does not, as developing countries originally desire, provide rules for financial compensation or remediation.

The climate regime is served by multiple institutions providing financial and intellectual resources. The Green Climate Fund (GCF) is the primary financial organ of both the UNFCCC and the Paris Agreement. The GCF is financed by member-parties to the UNFCCC. It has committed USD $5.2 billion to 111 projects covering both adaptation and mitigation.[25] The Global Environment Facility was previously the main financer of climate projects, but has now taken a secondary role to the more recent GCF.

The Intergovernmental Panel on Climate Change (IPCC) provides the science basis for international climate governance. It is an intergovernmental scientific process that builds a consensus-based depiction of the science of climate change (working group I), impacts (working group II) and mitigation (working group III). The IPCC provides both assessment reports every five years, as well as special reports both at its own discretion and at the request of the UNFCCC parties.

The IPCC is complemented by the United Nations Environment Programme (UNEP), which has provided an abundance of report on climate governance. These include rolling reports on the mitigation gap, adaptation gap and climate finance.

The proliferation of climate-related law and institutions had a watershed moment in 2015. The Paris Agreement was met with a raft of long-awaited climate-relevant policy announcements. These included the Kigali Amendment to phase out hydrofluorocarbons (HCFs, a potent greenhouse gas and replacement for ozone depleting substances), the Carbon Offsetting Scheme for International Aviation (CORSIA) Under the International Civil Aviation Authority (ICAO) and goal 13 of the Sustainable Development Goals. In 2018 the International Maritime Organisation's (IMO) Marine Environment Protection Committee (MEPC) released an initial strategy on emissions reductions from shipping. This includes an aim to peak emissions from shipping as soon as possible and reduce them by 50% by 2050 compared to 2008 levels. These different initiatives form a cluster of complementary mitigation efforts outside of the central climate regime. However, in terms of the efficacy of these initiatives, the SDGs are non-binding and offer no concrete targets or mechanisms. The CORSIA agreement is a voluntary agreement based on offsetting. The IMO strategy offers high-level, non-binding strategic guidance with goals that are not congruent with limiting warming to 2 °C.

Mitigation and adaptation activities are also carried out by a range of other intergovernmental bodies. Mini-lateral forums such as the G20, G8 and Major Economies Forum have all made multiple statements regarding climate change. These are non-binding political declarations but can help to mould norms and build political momentum.

Adaptation and mitigation actions are occurring through a range of UN agencies and affiliated institutions. These include large climate finance programmes from the World Bank, European Investment Bank (EIB), European Bank for Reconstruction and Development (EBRD). Numerous UN agencies, such as the United Nations Development Programme (UNDP) are looking to mainstream climate adaptation and mitigation considerations into their projects and programmes.

Actions by subnational and non-state actors are loosely linked to the climate regime. The "NAZCA" platform is a database of non-state and subnational climate actions and pledges maintained by the UNFCCC Secretariat. While it is a useful depository for tracking international efforts, it has no mandate for comparing, critiquing or influencing non-state actions. The actions of sub-national entities such as cities, localities and regions are undertaken through a range of networks including ICLEI (Local Governments for Sustainability, a network of more than 1,750 local and regional governments), C40 Cities for Climate Leadership and the Global Covenant of Mayors for Climate and Energy.

While mitigation and adaptation are well covered broadly, the response to tipping points or global catastrophe is not. Scientific knowledge of tipping points[26] and early warning signals[27] has progressed substantially. Yet the primary instruments of the climate regime do not have dedicated mechanisms to either induce a rapid response in the case of a looming tipping point, nor to adapt to or recover from an unforeseen climate catastrophe. International climate governance is focused on the average, rather than high-impact, low-probability "tail risks".

A second blind spot is supply side governance. Regulating the extraction, development and refining of fossil fuels offers numerous economic and political advantages.[28] Yet the Paris Agreement makes no mention of fossil fuels. None of the instruments of the climate regime ban the exploration or development of fossil fuels. This has led to recent calls for an international fossil fuel non-proliferation treaty.[29]

Importantly, the existing governance has not been successful in diverting the world away from dangerous warming. Current emissions trajectories have the world moving towards warming between 2.0–4.9 °C by 2100,[30] with a median of around 2.6–3.1 °C[31] or 3.1–3.5 °C.[32] The Paris

Agreement is unlikely to be able to bend the emissions curve down to 1.5–2 °C. Both weak compliance mechanisms, an unproven method of ″ratcheting up″ commitments, and the lock-in of emissions-intensive infrastructure by 2020 all undermine the effectiveness of the agreement.[33]

Table 5: Coverage and gaps in governance to prevent catastrophic or extreme climate change.

Coverage	Wide-reaching coverage of the science of climate change science, impacts and mitigation. Mitigation, adaptation, loss and damages, market-based mechanisms, are covered primarily by the UNFCCC-centred regime, and a raft of other initiatives.
Gaps	Governance of catastrophic or extreme climate change, response to tipping points and early warning signals, stranded assets, fossil fuel non-proliferation.
Issues requiring attention	All of the issues outlined above require critical attention. There is already some nascent research on fossil fuel non-proliferation. Research on catastrophic warming and the potential for "tail-risk treaties" are a neglected and high importance priority.

3.5 Solar geoengineering

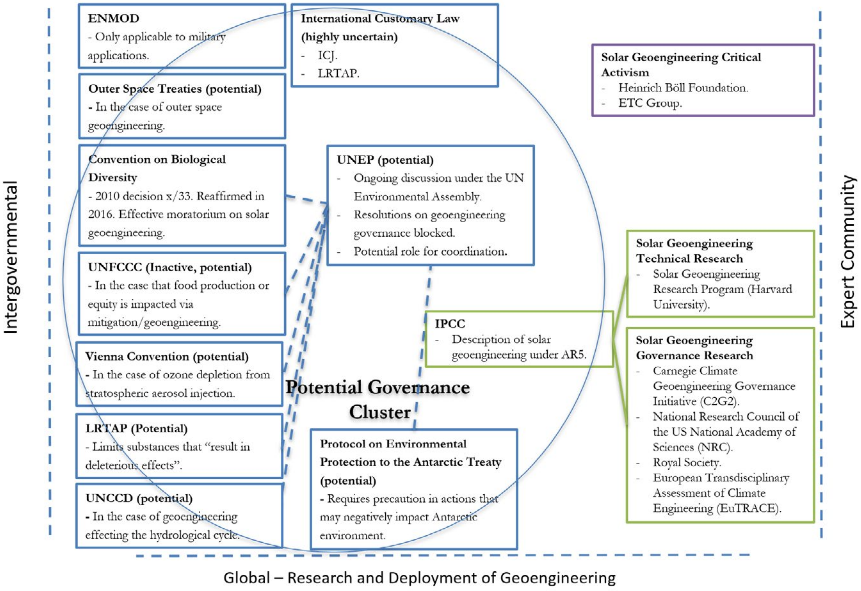

Intergovernmental
ENMOD
- Only applicable to military applications.
International Customary Law (highly uncertain)
- ICJ.
- LRTAP.
Solar Geoengineering Critical Activism
- Heinrich Böll Foundation.
- ETC Group.
Outer Space Treaties (potential)
- In the case of outer space geoengineering.
Convention on Biological Diversity
- 2010 decision x/33. Reaffirmed in 2016. Effective moratorium on solar geoengineering.
UNEP (potential)
- Ongoing discussion under the UN Environmental Assembly.
- Resolutions on geoengineering governance blocked.
- Potential role for coordination.
Expert Community
Solar Geoengineering Technical Research
- Solar Geoengineering Research Program (Harvard University).
UNFCCC (Inactive, potential)
- In the case that food production or equity is impacted via mitigation/geoengineering.
IPCC
- Description of solar geoengineering under AR5.
Solar Geoengineering Governance Research
- Carnegie Climate Geoengineering Governance Initiative (C2G2).
- National Research Council of the US National Academy of Sciences (NRC).
- Royal Society.
- European Transdisciplinary Assessment of Climate Engineering (EuTRACE).
Vienna Convention (potential)
- In the case of ozone depletion from stratospheric aerosol injection.
Potential Governance Cluster
LRTAP (Potential)
- Limits substances that "result in deleterious effects".
Protocol on Environmental Protection to the Antarctic Treaty (potential)
- Requires precaution in actions that may negatively impact Antarctic environment.
UNCCD (potential)
- In the case of geoengineering effecting the hydrological cycle.
Global – Research and Deployment of Geoengineering

Fig. 8: Solar engineering regime complex.

There is no explicit international governance of solar geoengineering. As shown in Figure 3, there is a large cluster of treaties which could be relevant. However, these are unplanned, incidental and piecemeal with limited ability for binding application.[34] Thus, there is widespread agreement that there is no distinct solar geoengineering regime and a need for direct governance.[35]

There that norms and rules around environmental impact assessments and harms from transboundary pollution have relevance in guiding the testing and use of such technologies. For example, the International Court of Justice has affirmed that states have a duty under international customary law to avoid major transboundary harm to either the global environmental commons or the territory of other states.[36] However, the application of customary international is highly uncertain and unlikely to be effective in overseeing or deterring unilateral or multilateral deployment of solar geoengineering, or even smaller field-tests.[37]

The most direct piece of solar geoengineering governance is the 1976 Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques ("ENMOD Convention"). ENMOD was established in the wake of US attempts to weaponise weather manipulation during the Vietnam War. It appears to have been successful in curtailing research efforts into weather modification. By 1979 US research into the area had declined sharply.[38] However, the use of ENMOD is limited for solar geoengineering as it only covers military applications. The preamble of ENMOD actively endorses the potential civilian uses of geoengineering type activities: "*... the use of environmental modification techniques for peaceful purposes could improve the interrelationship of man and nature and contribute to the preservation and improvement of the environment for the benefit of present and future generations*." Given that the majority of use cases of solar geoengineering are likely to be civilian, ENMOD is of restricted utility.

In lieu of any overarching authority, the Convention on Biological Diversity (CBD) has undertaken action on governing geoengineering research and deployment. In 2010 the CBD adopted a decision which could be taken as a de-facto moratorium on large-scale geoengineering. Paragraph (w) of decision x/33 states: "*that no climate-related geo-engineering activities that may affect biodiversity take place, until there is an adequate scientific basis on which to justify such activities and appropriate* *consideration of the associated risks for the environment and biodiversity and associated social, economic and cultural impacts*."[39] There is an exception for small-scale scientific research studies that can be performed in a controlled environment. The decision was reasserted in 2016, with the caveat that further transdisciplinary research and knowledge-sharing was needed to understand governance options and the potential impacts.[40] However, these are non-binding decisions, and ultimately the CBD lacks enforcement mechanisms. It also lacks the participation of one of the most credible potential developers of solar geoengineering: the US.

While international legal arrangements are sparse, there has been a groundswell of work from expert communities. A watershed moment was the 2009 Royal Society Report into governance and ethical issues. This was followed by a 2010 report examining geoengineering regulation by the UK House of Commons Scientific and Technology Committee, a 2011 report by the Kiel Earth Institute, a 2013 piece by the Congressional Research Service in the US and 2015 assessment by the European Transdisciplinary Assessment of Climate Engineering (EU-TRACE).[41] Geoengineering was then covered in the IPCC's fifth Assessment Report (AR5) in 2014 and will be investigated in further depth in AR6.

Geoengineering governance is now a well-established sub-field with academics across multiple institution involved. Technical research has been slower due to social concerns and the previous failure of the 2011 SPICE (Stratospheric Particle Injection for Climate Engineering) programme.[42] The experiment has sought to field-test a delivery system for stratospheric aerosol injection but faced severe public backlash.

Table 6: Coverage and gaps in governance of solar geoengineering.

Coverage	Existing governance arrangements are limited to moratoriums and work programmes under different bodies.
Gaps	Almost all SRM activities are not covered under any form of binding international law. This includes rules for deployment, maintenance, innovation or research into the science of geoengineering.
Issues requiring attention	The governance of solar radiation management and the unilateral or plurilateral deployment of stratospheric aerosol injection.

3.6 Unknown risks

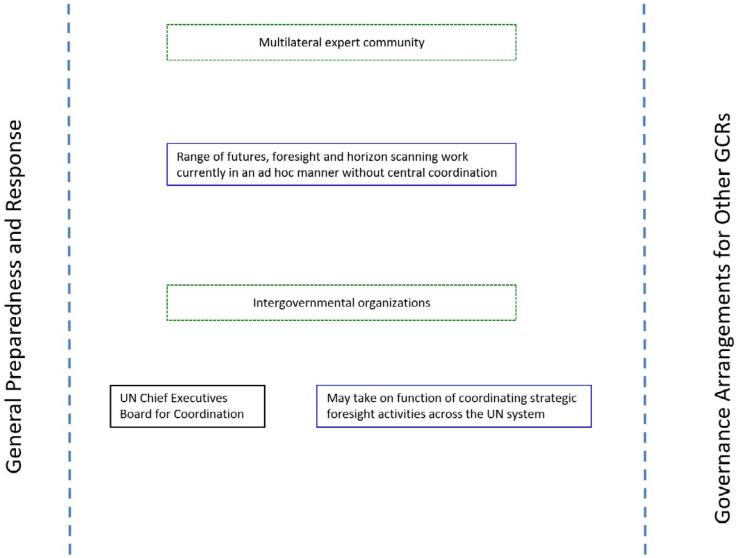

General Preparedness and Response
Multilateral expert community
Range of futures, foresight and horizon scanning work currently in an ad hoc manner without central coordination
Intergovernmental organizations
UN Chief Executives Board for Coordination
May take on function of coordinating strategic foresight activities across the UN system
Governance Arrangements for Other GCRs

Fig. 9: Unknown risks regime complex.

There are two particular elements to consider when seeking to identify global governance arrangements for unknown GCRs:

- Processes that might help with identification and analysis of emerging threats.
- General governance arrangements relating to preparedness, resilience and response to GCRs.

And, while we do not know the source and mechanism of unknown risks, we do have some knowledge about the likely objects of protection — that is, what it is we seek to protect from any such risk — and therefore which areas of governance we might look to for developing responses should such risks become apparent. For example, whatever the source of risk, we are likely to be interested in protecting human, animal and plant life and health and stability of planetary life support systems.[43]

General GCR governance arrangements are addressed in Section 5.

Processes that might help with identification and analysis of emerging threats involve futures studies, foresight and horizon-scanning work, and a range of approaches are available within this. Such activities do have some limitations, and using a combination of approaches, and joining up different exercises can address some of these. They necessarily face their greatest limitations in identifying unknown risks, but there are some techniques for approaching this: for example, through use of "wild cards". Involving a wide range of expertise within such processes will also have greater value, particularly because unknown risks may be more likely to occur at the intersection of e.g. different technological areas.

Some national governments and agencies within them regularly undertake foresight activities, often with an aim of identifying potential emerging threats. Such activities are also undertaken by research institutions, science academies, and professional organisations. Global governance of unknown risks could benefit from mechanisms to bring together information from such exercises, so that analysis can be conducted over time and across different countries, regions and sectors.

Several of the international organisations involved in GCR governance conduct simulation exercises which can serve a similar function by helping to identify potential gaps and challenges in responding to emerging threats, and some are exploring the potential use of foresight activities for their work (although not necessarily with the aim of identifying unknown risks, such processes might be adapted to do so). Science and technology review processes associated with some of the organisations may also be a useful basis for such work (although they tend to focus on shorter-term horizons or recent developments). Existing systems for early warning and surveillance may also help to identify novel threats: for example, the health impacts of a novel risk may be picked up before the source of the risk is identified.

The UN Chief Executives Board for Coordination, which is made up of the heads of the UN's specialised agencies, funds and programmes, and focuses on fostering coordination and coherence across the UN system, is examining the opportunities for "integrating strategic foresight into its work and... for promoting foresight capacities and fostering collaboration across the system"[44] — if pursued such activities could significantly enhance foresight capacities at the global level.

Table 7: Coverage and gaps in preparedness for unknown GCRs.

Coverage	There are some formative but no well-established global governance arrangements for identifying unknown GCRs. Some current surveillance and monitoring systems might support detection of unknown risks. Preparedness, resilience and response will depend on more general GCR governance arrangements.
Gaps	There are gaps across this area. Given the inherent difficulties in identifying and detecting unknown risks, while efforts for this should not be neglected, enhancing general capabilities in preparedness, resilience and response to GCRs would seem to be a higher priority (particularly because of the benefits this would bring to known GCRs governance too).
Issues requiring attention	Development and implementation of robust foresight activities for identifying emerging threats. Research on how to enhance the ability of existing surveillance and monitoring systems to spot indications of unknown risks, and communicate with relevant communities to investigate them. Increasing the capabilities of broader GCRs governance.

3.7 Nuclear warfare

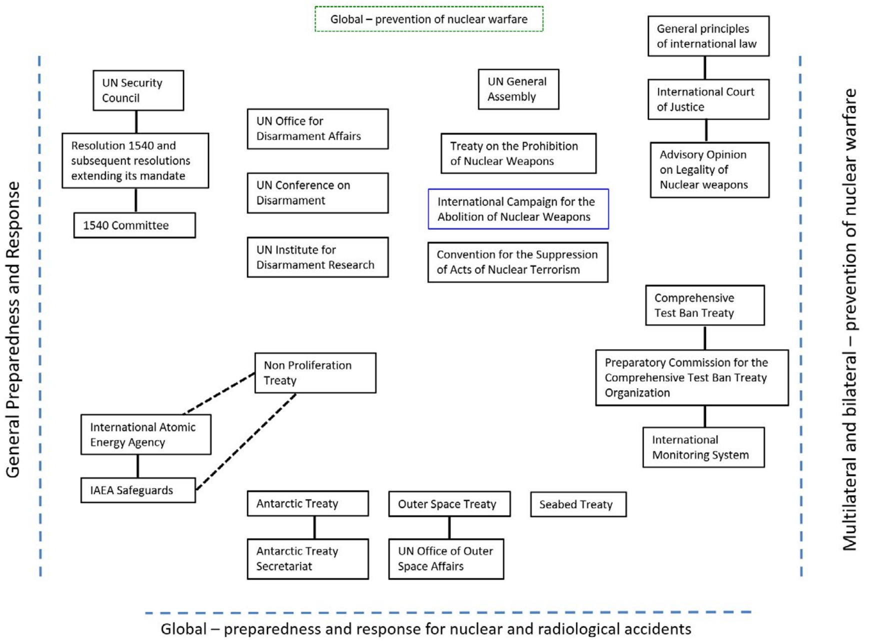

Global – prevention of nuclear warfare
General principles of international law
International Court of Justice
Advisory Opinion on Legality of Nuclear weapons
UN Security Council
Resolution 1540 and subsequent resolutions extending its mandate
1540 Committee
UN Office for Disarmament Affairs
UN Conference on Disarmament
UN Institute for Disarmament Research
UN General Assembly
Treaty on the Prohibition of Nuclear Weapons
International Campaign for the Abolition of Nuclear Weapons
Convention for the Suppression of Acts of Nuclear Terrorism
Comprehensive Test Ban Treaty
Preparatory Commission for the Comprehensive Test Ban Treaty Organization
International Monitoring System
Non Proliferation Treaty
International Atomic Energy Agency
IAEA Safeguards
Antarctic Treaty
Antarctic Treaty Secretariat
Outer Space Treaty
UN Office of Outer Space Affairs
Seabed Treaty
General Preparedness and Response
Multilateral and bilateral – prevention of nuclear warfare
Global – preparedness and response for nuclear and radiological accidents

Fig. 10: Nuclear warfare regime complex.

The extremely severe effects of an all-out nuclear war mean that prevention of such an event is a priority for global governance efforts. If such an event were to occur a response to a range of damage would be needed. Large numbers of immediate fatalities would be expected, particularly where major population centres are targeted (80–95% in a 1–4 km radius);[45] there would be significant health impacts on a large scale, compounded by the loss of health systems, staff and infrastructure; widespread environmental contamination; extensive disruption to critical infrastructure; large-scale migration; and probably continued geopolitical instability. If at a sufficient scale to cause "nuclear winter" the associated collapse in global agricultural production would result in global famine and starvation.

Global governance specific to nuclear warfare focuses on:

- Reductions in armaments, with an eventual goal of general and complete nuclear disarmament.
- Preventing proliferation of nuclear weapons and diversion of nuclear materials.
- Measures to stabilise relations between nuclear states, avoid misinterpretation through communication mechanisms, and build confidence through verification and inspection arrangements.
- Creation of nuclear-weapon free zones.

Legal arrangements include:

- Some global legal instruments including a general prohibition on nuclear weapons. These instruments have not all achieved participation of (all) nuclear weapons states, and some are yet to enter into force.
- Some multilateral agreements, generally around creation of nuclear-weapon-free zones, and involving a set of regional states and accompanied by protocols that commit nuclear weapons states to not testing or using nuclear weapons within those zones.
- Some bilateral agreements, primarily between the US and Russia.

The main international organisations operating in this area include the UN Security Council and General Assembly, Conference on Disarmament, and Office for Disarmament Affairs (UNODA), the Preparatory Committee for the Comprehensive Test Ban Treaty Organization (CTBTO), and International Atomic Energy Agency (IAEA). There are also some multilateral export control groups.

(Specific details on each of the legal instruments, agreements and organisations is provided in Appendix I, available online).

Civil society movements have played a significant role in shaping global governance of nuclear warfare, particularly through: the International Campaign to Abolish Nuclear Weapons (ICANW), which was pivotal in bringing about the successful negotiation of the Nuclear Weapons Prohibition Treaty; the World Court Project that prompted states to take the issue of legality of nuclear weapons to the International Court of Justice; and in establishing nuclear-weapon-free cities, local authorities, and regions. Expert networks support the work of the IAEA, and the Preparatory Committee of the CTBTO, particularly its monitoring systems.

The humanitarian impacts of nuclear war, including nuclear winter scenarios, have motivated a lot of these global governance efforts; however, addressing such impacts is largely outside the focus of these legal arrangements (aside from some generalised commitments to assist states attacked with nuclear weapons).

Some international organisations' work, which relates to dealing with nuclear accidents, may provide a basis for such responses, but it is generally unclear whether this would be possible and how adaptable and scalable such activities might be. This work includes: two IAEA conventions; guidance documents; and networks of emergency responders and other experts. The effectiveness of a response is therefore likely to depend on general global governance for disaster preparedness, resilience and response and emergency management. This is unlikely to be systematic enough to deal with the full range of immediate through to long-term impacts, nor adequate for such a scale of catastrophic event. Such capacities may also have been damaged or impeded by the geopolitical instability that resulted in nuclear war.

Table 8: Coverage and gaps in governance of nuclear risks.

Coverage	There are a large number of international legal instruments that address various aspects of the prevention of nuclear warfare. These have had some notable success in reducing armaments, but not yet to a level which is likely to avoid nuclear winter scenarios in all-out nuclear war. There has also been some success in limiting proliferation of nuclear weapons capabilities, though this is regularly challenged. There is a strong international norm against testing, use and possession of nuclear weapons, but no indication that nuclear weapons states will make significant moves towards disarmament in the coming decades, and some indications that the US in particular is moving away from armament restraint. The IAEA does extensive work promoting nuclear safety and security and checking safeguards that back up the Non-Proliferation Treaty, and the Preparatory Committee for the CTBTO is establishing a robust monitoring network. There is very limited coverage of preparedness and response for the impacts of a large-scale nuclear wear; this will largely depend on more general global governance arrangements.
Gaps	The most significant gaps are likely to be found in the general global governance of disaster preparedness, resilience and response.
Issues requiring attention	Continued pressure needs to be applied to nuclear weapons states to further reduce their arsenals, and subsequently to comply with the international prohibition on nuclear weapons. Within this careful attention will be needed to the stability of deterrence. There continue to be a few states that fail to comply with non-proliferation arrangements, and / or that express a desire to attain nuclear weapons. As noted in GCF's *Global Catastrophic Risks 2018 Report*, continued efforts to address regional conflict and geopolitical instability are important.

3.8 Super-volcanic eruption

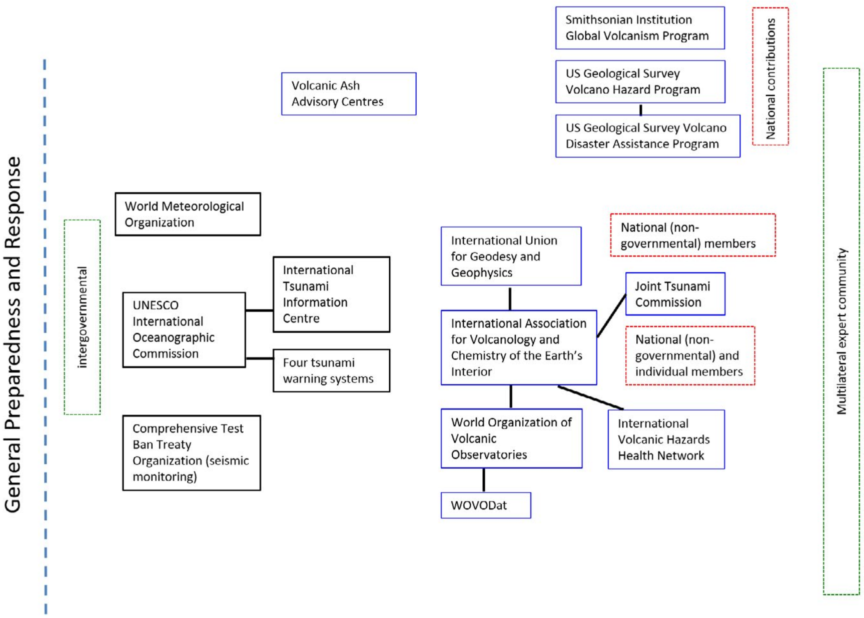

General Preparedness and Response
intergovernmental
World Meteorological Organization
UNESCO International Oceanographic Commission
International Tsunami Information Centre
Four tsunami warning systems
Comprehensive Test Ban Treaty Organization (seismic monitoring)
Volcanic Ash Advisory Centres
Smithsonian Institution Global Volcanism Program
US Geological Survey Volcano Hazard Program
US Geological Survey Volcano Disaster Assistance Program
National contributions
International Union for Geodesy and Geophysics
International Association for Volcanology and Chemistry of the Earth's Interior
National (non-governmental) members
Joint Tsunami Commission
National (non-governmental) and individual members
World Organization of Volcanic Observatories
International Volcanic Hazards Health Network
WOVODat
Multilateral expert community

Fig. 11: Super-volcanic eruption regime complex.

Global governance arrangements specific to super-volcanic eruptions are sparse and primarily limited to expert networks and collaborating research institutions. These mainly focus on scientific and technical aspects of monitoring and observation. Given limited (if any) prevention capability, most of the activities addressing impacts will fall under general global governance of disaster preparedness, resilience and response. The areas that need to be addressed include: immediate impacts including large-scale loss of life and damage to critical infrastructure (some volcanoes are in areas with local populations of over five million, and there could also be resulting tsunamis effecting other regions); and the longer-term global impacts associated with climate disruptions and resulting agricultural production losses, which could result in widespread starvation.

Super-volcanic eruptions are predicted to occur far more frequently than globally catastrophic asteroid impacts (~1 in 17,000 years compared to ~1 in several hundred thousand years), so the even more limited governance response probably represents a major gap, particularly if the general global governance of disaster preparedness, resilience and response is inadequate. Anyway, improved global coordination of the research, observation, monitoring and early warning of volcanic eruptions would be beneficial.

Central to international coordination efforts is the International Association of Volcanology and Chemistry of the Earth's Interior (IAVCEI, an association of the International Union of Geodesy and Geophysics). National members of the International Union of Geodesy and Geophysics (IUGG), of which there are 72 currently, and its associations participate in a non-governmental capacity. 52 IUGG national members participate in of IAVCEI, which also has individual members. Two commissions of the IAVCEI have particular relevance to governance of super-volcanic eruptions: the World Organization of Volcano Observatories (WOVO), which facilitates cooperation between 80 observatories located in 33 countries; and the International Volcanic Health Hazards Network, an interdisciplinary expert network, which collates research and disseminates information on volcanic health hazards and impacts.

A lot of data collection, analysis and dissemination is done by national-based bodies and research institutions such as the Smithsonian Institution's Global Volcanism Program and the US Geological Survey's Volcano Hazard Program. The latter includes a Volcano Disaster Assistance Program, which provides expert support and equipment during volcano crisis events worldwide. WOVODat (linked to the WOVO and hosted by the Earth Institute of Singapore) is also building a global database on volcanic unrest with the aim of improving eruption prediction. Data from seismic monitoring conducted as part of the activities of the Comprehensive Test Ban Treaty Organization may also contribute to data on volcanic unrest.

Nine Volcano Ash Advisory Centres serve the needs of the aviation industry during eruption events. The World Meteorological Organization also provides advice for aviation following eruptions, and may provide information about weather and climate impacts of eruptions. It also has a general disaster risk reduction programme which includes impacts from volcanic eruptions and tsunamis.

Depending on the eruption site, a tsunami may follow a super-volcanic eruption. IAVCEI along with two other IUGG Associations (the International Association of Seismology and Physics of the Earth's Interior, and the International Association of Meteorology and Atmospheric Sciences) have a Joint Tsunami Commission to exchange scientific and technical information with countries that may be affected by tsunamis. The United Nations Educational, Scientific and Cultural Organization (UNESCO)'s International Oceanographic Commission (IOC) has a Tsunami Programme which includes intergovernmental committees for four warning systems (covering the Pacific, Indian Ocean, Caribbean, and North East Atlantic and Mediterranean). The IOC has a mandate to develop a global Tsunami warning system, but this has not yet been established. The International Tsunami Information Centre, associated with the Pacific Tsunami Warning and Mitigation System also carries out programmes for risk assessment and for local community education on preparedness.

UNESCO also has a Geohazards Programme, focusing on associated disaster risk reduction, management and mitigation. While its overview mentions super-volcanic eruptions as events that can threaten humankind, but there doesn't seem to be any work addressing them in the programme.

Table 9: Coverage and gaps in governance for super-volcanic eruptions.

Coverage	Extremely limited with a dominant focus on information sharing and research collaboration for observation, monitoring and early-warning. It will largely demand on general global governance arrangements for disaster preparedness, resilience and response.
Gaps	There are significant gaps in global governance for super-volcanic eruptions. There is no global organisation with a mandate to manage volcanic risk, and no standardised international system for volcano alert levels. Not all sites are adequately monitored. There is little indication of establishment of global norms, e.g. around benefit to humanity. Given limited warning time and no means of prevention of super-volcanic eruptions, substantial attention to preparedness, resilience and response is needed. General disaster governance efforts are unlikely to be adequate.
Issues requiring attention	This is a generally neglected area and underfunded area. Sustainability and continuity particularly of non-intergovernmental arrangements. There do not appear to be the same norms around openness and data sharing that there are for near-Earth objects, for example WOVODat has a two-year grace period for release of new data.

4. Drivers and Vulnerabilities

Global Catastrophic Risk is not just a reflection of hazards, but also underlying vulnerabilities. The governance of these vulnerabilities is just as crucial as that of hazards. Yet the coverage of vulnerabilities, both in nature and governance, is far less developed. As a starting point, we will draw on a listing of different contributors to the collapse of previous civilisations.[46] Given that previous societal collapses are the closest recurring analogues we have to GCRs, this is a prudent step. The collapse contributors include environmental degradation, climatic change, declining returns on complexity, declining returns on energy, inequality, oligarchy, as well as external shocks such as disease, warfare and natural disasters. Many of these have already been covered in our cartography, including climate change, (nuclear) warfare, disease, and GCR relevant natural disasters. Others, such as complexity and returns on energy investment, are too nascent and theoretical to be approached directly. This leaves us with environmental degradation, inequality and oligarchy. We will subsume oligarchy under inequality and address these as the two fundamental drivers of GCRs to be examined.

4.1 Inequality

Wealth inequality tends to increase inexorably over time[47] and has been linked to both historical societal collapses,[48] as well as other catastrophes such as world wars.[49] This section will explore inequality both in terms of wealth and income inequality within and between countries. There are two separate forms of governance covering these areas:

- Equality-inducing measures within international treaties;
- Governance of mechanisms that drive inequality, primarily tax avoidance and evasion.

The significant global arrangements for poverty alleviation could also be considered as part of international efforts to reduce inequality. This includes both Official Development Assistance (ODA) guidelines under the OECD and the international financial institutions involved with economic development, such as the World Bank and International Monetary Fund. However, the explicit goal of these efforts and infrastructure is poverty alleviation and economic development, not the alleviation of poverty. To the contrary, efforts such as structural adjustment programmes likely worsened inequality both within numerous developing countries and the between developed and developing countries.[50] Instead, we will focus on mechanisms for equity across treaties, and the governance of tax evasion and avoidance.

There is no explicit international governance of income or wealth or income inequality. The closest shadow of direct governance is SDG 10 "Reduce inequality within and among countries".[51] While the headline is compelling, the targets are ambiguous and do not set any concrete objectives or measures. Moreover, the SDGs are a non-binding declaration lacking any credible mechanism for ensuring compliance.

Equity considerations are split across multiple treaties and bodies. It has been integral to most environmental treaties. The CBD, UNFCCC and most other multilateral environmental agreements contain numerous capacity-building measures as well as financial support provisions for developing countries. This is underpinned by the principles of environmental multilateralism as enshrined in the 1992 Rio Declaration on Environment and Development. Principle 3 states the development must "equitably meet developmental and environmental needs of present and future generations". Principle 5 notes the need for poverty eradication to avoid major international disparities and principle 6 notes that special priority should be given to developing countries. Mechanisms for capacity building and financial transfer are not just common across environmental agreements, but also in the areas of trade, health and security.

The international system also has a dedicated system to manage drivers of inequality such as tax evasion and avoidance. The United Nations Conference on Trade and Development (UNCTAD), OECD, G20 and IMF all provide estimates of tax evasion both as revenue base erosion and profit sharing. Both the Global Forum on Transparency and Exchange of Information for Tax Purposes and the Multilateral

Convention on Mutual Administrative Assistance in Tax Matters (MCMAATM) provide a platform for the exchange of basic tax information.[52] The framework is unlikely to stem tax evasion or global inequality until more drastic measures, such as global wealth tax, are introduced.[53]

The existing framework to tackle the problem appears to be inadequate. By most measures, global inequality is deteriorating. The typical measurement of the Gini index suggests that inequality between countries has decreased over the past decade over half. In 2000 it was approximately 44, but had dropped to 39 by 2016.[54] This is largely due to the economic rise of major developing countries such as China and India. However, the inequality measured by the Gini index has worsened within most countries over the past few decades, particularly OECD countries.[55] Other measurements portray an even worse situation. The global wealth share of the top 1% has grown from 25–30% in the 1980s to roughly 40% in 2016.[56] The real figure is likely to be far worse once the hidden treasures of tax havens are considered.[57] This is testament to the ineffectiveness of the existing patchwork that governs inequality internationally.

Table 10: Coverage and gaps of global wealth inequality.

Coverage	Wealth and income inequality is partially covered by the OECD centred tax regime. This has largely been unsuccessful in addressing the dynamics which exacerbate wealth inequality such as tax evasion and regressive taxation. Many treaties and bodies contain provisions for capacity building and equity, but their success is dubious.
Gaps	The governance of wealth inequality between and within countries is largely a glaring gap in international arrangements.
Issues requiring attention	The abolition of channels to inequality, such as tax evasion, as well as mechanisms to mitigate and reverse wealth and income inequality between and within countries.

4.2 Ecological collapse

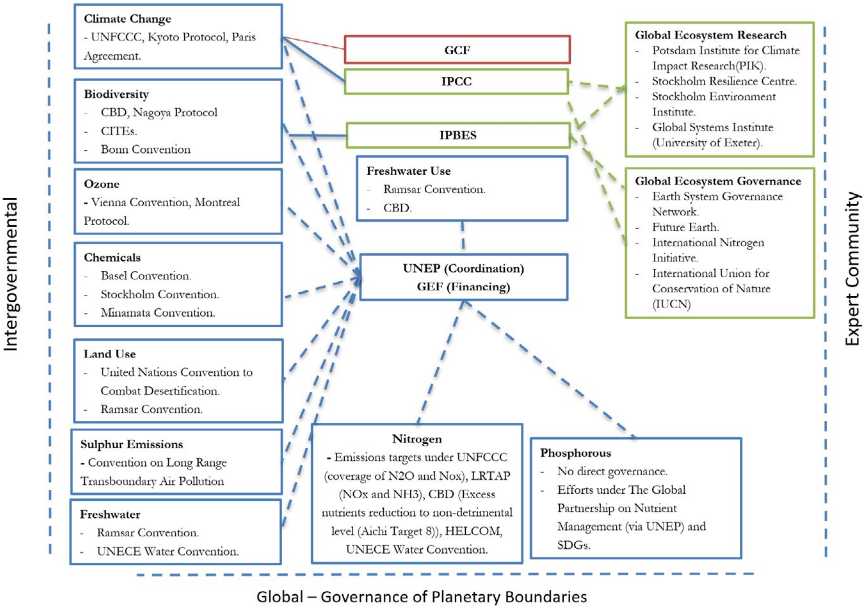

Climate Change
- UNFCCC, Kyoto Protocol, Paris Agreement.
Biodiversity
- CBD, Nagoya Protocol
- CITEs.
- Bonn Convention
Ozone
- Vienna Convention, Montreal Protocol.
Chemicals
- Basel Convention.
- Stockholm Convention.
- Minamata Convention.
Land Use
- United Nations Convention to Combat Desertification.
- Ramsar Convention.
Sulphur Emissions
- Convention on Long Range Transboundary Air Pollution
Freshwater
- Ramsar Convention.
- UNECE Water Convention.
GCF
IPCC
IPBES
Freshwater Use
- Ramsar Convention.
- CBD.
UNEP (Coordination)
GEF (Financing)
Nitrogen
- Emissions targets under UNFCCC (coverage of N2O and Nox), LRTAP (NOx and NH3), CBD (Excess nutrients reduction to non-detrimental level (Aichi Target 8)), HELCOM, UNECE Water Convention.
Phosphorous
- No direct governance.
- Efforts under The Global Partnership on Nutrient Management (via UNEP) and SDGs.
Global Ecosystem Research
- Potsdam Institute for Climate Impact Research(PIK).
- Stockholm Resilience Centre.
- Stockholm Environment Institute.
- Global Systems Institute (University of Exeter).
Global Ecosystem Governance
- Earth System Governance Network.
- Future Earth.
- International Nitrogen Initiative.
- International Union for Conservation of Nature (IUCN)
Intergovernmental
Expert Community
Global – Governance of Planetary Boundaries

Fig. 12: Planetary boundaries regime complex.

The loss of ecosystem services is a loss of humanity's resilience. Ecological is a broad phenomenon and it is difficult to draw clear contours around it. We will use the planetary boundaries framework to focus our analysis. The framework puts forward nine key global environmental services that constitute a "safe operating space for humanity": climate change, biodiversity loss, the nitrogen cycle, phosphorous, ocean acidification, land use, freshwater, ozone depletion, atmospheric aerosols and chemical pollution.[58]

It would be impossible to depict and analyse all of the agreements relevant to the governance of ecological collapse. The International Environmental Agreements Database lists "1,300 multilateral environmental agreements (MEAs), over 2,200 bilateral environmental agreements (BEAs), 250 other environmental agreements, and over 90,000 individual country 'membership actions'".[59] Instead, we will examine the primary instruments governing each planetary boundary. Climate change will be excluded from the analysis, as it has already been investigated.

As shown in Figure 4, the governance of ecological collapse is fragmented institutions. While the diagram depicts governance clusters for each of the planetary boundaries, this is not the case for many. Governance of land-use, freshwater, nitrogen and phosphorous are all deeply fragmented with no overarching convention of framework.

The United Nations Environment Programme (UNEP) acts as the coordinator of UN's multilateral environmental agreements. In practice, it has struggled to ensure effective collaboration and action between the multitude of agreements.

Most of the governance arrangements are served by the Global Environment Facility (GEF). The GEF acts as the primary financier of international environmental governance. It is financially replenished every four years by its 39 donor country members.[60] It covers forests, international waters, biodiversity, climate change, land degradation, chemicals and other areas, using a variety of grant and non-grant financial instruments.[61]

- *Biodiversity loss*: Biodiversity loss is directly governed by the 1992 Convention on Biological Diversity (CBD) and its 2010 Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising From Their Utilization to the Convention on Biological Diversity. The CBD provides rules and guidance on biodiversity monitoring and reporting, management actions and targets to reduce biodiversity loss. It is scientifically served by the Intergovernmental Panel on Biodiversity and Ecosystem Services (IPBES). The biodiversity regime is complemented by the trade-focused 1973 Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) and the 1979 Convention on the Conservation of Migratory Species of Wild Animals (1979 Bonn Convention).
- *Chemical pollution*: The international governance of chemical pollution centres upon a trio of treaties: the 2001 Stockholm Convention on Persistent Organic Pollutants; the 1989 Basel Convention on the Control of Transboundary Movements of Hazardous Wastes and Their Disposal; and the 2017 Minamata Convention on Mercury.

- *Ozone*: The international regime is the posterchild for effective environmental multilateralism. It is underpinned by the 1985 The Vienna Convention for the Protection of the Ozone Layer and the 1987 Montreal Protocol on Substances That Deplete the Ozone Layer. The treaties have been successful in addressing the problem of ozone depletion. The use of ozone depleting substances has been decreasing over the past two decades. Recent satellite data suggests that the hole in the ozone layer is now beginning to shrink and recover. This is largely due to the Montreal Protocols strong non-party mechanism (restricting trade in ozone depleting substances with non-parties) and enforcement mechanism.
- *Atmospheric aerosols*: The 1979 Convention on Long-Range Transboundary Air Pollution (LRTAP) is the primary international instrument for regulating sulphur emissions.
- *Nitrogen*: There is no explicit governance framework for nitrogen. Instead, there are targets relevant to nitrogen usage split across multiple multilateral environmental agreements. These include emissions targets under the UNFCCC (coverage of N2O and Nox), LRTAP (NOx and NH3), CBD (excess nutrients reduction to non-detrimental level (Aichi Target 8)), HELCOM (Baltic Marine Environment Protection Commission — Helsinki Commission), the OSPAR Commission (Nitrogen oxides or their transboundary fluxes impacting eutrophication), United Nations Economic Commission for Europe (UNECE) Water Convention (nitrate and nitrite concentrations).[62] The diversity and fragmentation appears to have hindered efforts. A more integrated regime that targets nitrogen pollution origins would be preferable.[63]
- *Phosphorous*: Like nitrogen, the governance of phosphorous is fragmented. There are also non-legal approaches to governing both nitrogen and phosphorous. Foremost is the Global Partnership on Nutrient Management under UNEP.
- *Land-use*: Land-use governance occurs primarily through a duo of legal instruments: the 1994 UN Convention to Combat Desertification and the 1971 Ramsar Convention on Wetlands

of International Importance. The Ramsar Convention is largely a pledge and review system, requiring countries to voluntarily submit wetland areas of importance to be regulated under it. Both lack effective enforcement and compliance mechanisms.

- *Freshwater*: As with phosphorous and nitrogen, there is no framework convention or overarching legal instrument to govern freshwater usage and pollution. It occurs through a patchwork including the UNECE Water Convention and the Ramsar Convention.

While not part of the Planetary Boundaries framework, population growth is a key driver of our collective environment impact. Direct governance of population is almost non-existent. The United Nations Department of Economic and Social Affairs (UN DESA) contains a population division. Its role is relegated to demographic research, including population projections and analysis. It has no role in attempting to curb global population growth. Nor does any other UN legal instrument or framework.

Overall, the global governance of phosphorous, nitrogen, atmospheric aerosols, and freshwater are the largest gaps in the protection of planetary boundaries. However, other important oversights exist. There is little effective, coherent governance across boundaries given their deeply interconnected nature. The role of coordination largely falls to UNEP. However, as an under-resourced programme under the UN General Assembly it has often struggled to effectively fulfil this task. As with climate change, there are no mechanisms to govern tipping-points, early-warning signals and the aversion of catastrophic ecological collapses. While there is the general international norm of the "precautionary principle" its exact meaning and implementation has often been hindered by ambiguity.

Like inequality, indicators for ecosystem collapse have been worsening over time. Ecological footprint *per capita* has trended steeply upwards since 1960, as far back as records go.[64] The Living Planet Index, a composite measurement for biodiversity, has also been more than halved from 1970 to the present day.[65] This warning signals suggest that despite that the governance of planetary boundaries while abundant is porous and inadequate.

Table 11: Coverage and gaps in governance of planetary boundaries and ecological collapse

Coverage	Adaptation and mitigation of ozone-depleting substances are effectively governed by the Vienna Convention and Montreal Protocol. Biodiversity loss and climate change are well, but ineffectively, covered by the UNFCCC and CBD regimes. Chemical pollution is partially covered by a cluster of treaties including the Stockholm and Basel Conventions.
Gaps	The governance of most planetary boundaries is currently fragmented and focused on mitigation, adaptation and science. The governance of phosphorous, nitrogen, atmospheric aerosols, and freshwater are all largely neglected. There is little to no governance of catastrophic tipping points, or interactions between earth systems.
Issues requiring attention	Tail risk treaties across environmental issues, as well as early warning and tipping point responses both within and across planetary boundaries.

5. The Broader GCR Governance Landscape

5.1 UN governance

The UN contains broader governance arrangements that are relevant for GCRs. First and foremost is the UN Office for Disaster Risk Reduction (UNDRR), which oversees the implementation of the International Strategy for Disaster Reduction. This includes efforts to build resilience, coordinate emergency responses to disasters and ensure effective recovery. The Sendai Framework for Disaster Risk Reduction 2015–2030 was endorsed by the UN General Assembly in 2015 and provides four priorities and seven targets for action. However, both UNDRR and the Sendai framework are focused on non-GCR, natural hazards. Their efficacy and mandate in reducing GCRs is questionable. It was preceded by the Hyogo Framework for Action, which covered disaster risk reduction guidance for the decade of 2005–2015.

Disaster management splintered across a wide range of bodies including WMO and WHO. The WHO includes decisions and frameworks for disease outbreaks, risks in emergencies, poisoning, displaced peoples, complex emergencies (caused by warfare or the large-scale movement of people) and other areas. The United Nations Educational, Scientific and Cultural Organisation (UNESCO) contains numerous programmes to assist countries in reducing both climate and disaster risk. These include activities on geohazard risk reduction, water hazard risk reduction, school safety, tsunamis, disaster risk reduction in UNESCO designated sites and crisis management and post-crisis transitions. Actions in these areas focus on knowledge provision and capacity building.

There has been nascent, unsuccessful discussion of introducing intergenerational governance mechanisms into the UN. This includes the 2012 push for an Ombudsman for Future Generations (to be located under the Secretary General) at the Rio+20 negotiations and the Secretary General's 2013 report on "Intergenerational Solidarity and the Needs of Future Generations". The former was unsuccessful and the latter is a non-binding review. Successfully introduced mechanisms for intergenerational governance in the UN could have profound implications for GCR management and foresight under the UN.

Table 12: Coverage and gaps in governance of catastrophic risk by the United Nations.

Coverage	The UN broadly covers disaster risk through the UNDRR and Sendai framework. While these incorporate preparedness, emergency response and risk reduction, they are primarily focused towards natural disasters. The UN Security Council has the mandate to cover risk reduction and response for conflict-based risks.
Gaps	Foresight of GCRs and existential risks, as well as preparedness, response and recovery to worst-case scenarios, and risk reduction and response for and across anthropogenic GCRs are all lacking.
Issues requiring attention	Mechanisms to coordinate foresight, recovery, response and reduction of GCR, particularly anthropogenic risks.

5.2 Transnational governance

Taking an appropriately broad perspective on what is encompassed by global governance (as outlined in Section 2), there are various actors and activities beyond formal inter-governmental arrangements. Those significant for the global governance of GCRs include:

5.2.1 Individual experts and communities of expertise

For all GCRs a key need is for greater understanding about the risks and prevention, mitigation and response options. There is, therefore, a substantial need for contributions from a range of experts to address these areas. While some international organisations and treaty processes — and national delegations engaging with them — have some in-house expertise, this is not always the case and may well be insufficient, particularly when it comes to more extreme risk scenarios.

In some GCR governance regimes experts are quite well integrated in inter-governmental arrangements at various levels of formality (for example, in protection of human, animal and plant health). In others, clear spaces have formed in which expert communities play a key support role and help to address gaps (for example, in science and technology reviews associated with the Biological and Toxin Weapons Convention). In yet other regimes, inter-governmental activity is very limited and expert communities form the core of global governance efforts (for example, in the area of super-volcanic eruptions).

Expertise may be provided on an individual basis or collectively through a representative organisation (such as a scientific academy) or through participation in collaborative networks (such as laboratory networks supporting the work of the World Health Organization).

An extensive range of disciplinary and practical experience and expertise is needed for effective governance of GCRs. Careful consideration of how to bring knowledge together across fields and integrate it in governance activities is needed, and there is substantial scope for further research and practical action in this regard. This needs to be worked out — and exercised — well in advance of potentially catastrophic events otherwise interventions are more likely to fail. (For example, the lack of integration of social science in international responses to the 2014 Ebola outbreak has been recognised as a key failure point).[66]

The role of experts in global governance is not unproblematic. There need to be ways of assuring quality, relevance and legitimacy of expertise — which can be assisted, for example, by use of peer networks. Setting particular standards for qualifications and level of experience can be useful, but can also privilege participation by certain groups and limit representativeness. Transparency about potential conflicts of interest is also important. Sometimes relationships between expert communities and formal governance processes are difficult; as at other levels, international policy making is not always evidence-based and policy-makers can have unrealistic expectations about expert input, e.g. expecting a level of certainty that is not achievable. Resourcing of expert communities can also present challenges; being transparent about funding sources is important, and political difficulties could arise where one particular state or agency is the main source of support for a group. Some expert groups will be disadvantaged by lower levels of funding and access to other resources such as facilities, equipment or data. As identified in some of the regime summaries, there is also a need to be alert to the sustainability of governance efforts where they rely heavily on expert activities and to have contingency plans should a key funding source be withdrawn.

5.3.2 Civil Society Organisations (CSOs)

CSOs also perform valuable roles in relation to GCRs governance, some of which we highlight here (with further examples provided in some of the regime summaries):

- Form the basis for transformative global campaigns to address particular GCRs — the role of the International Campaign for the Abolition of Nuclear Weapons in advancing negotiations towards the 2017 Treaty on the Prohibition of Nuclear Weapons is a prominent recent example.
- Provide a route of connection from the local to global levels both in bringing citizens' concerns to the attention of international bodies and in connecting international governance initiatives back to local action. (For example, this can be seen in the connection between local communities and the work of the UN Office for Disaster Risk Reduction in the Community Practitioners Platform for Resilience).[67]
- Provide global connectivity between groups with aligned interests and concerns, amplifying their ability to effect action transnationally. Examples include the Global Fossil Fuels Divestment Movement,[68] and the Mayors for Peace initiative, which brings together over 7,800 cities worldwide to engage citizens in pursuit of nuclear disarmament.[69]

5.3.3 Industry organisations and transnational corporations

Companies also have a significant role to play in global governance. This is frequently perceived / portrayed negatively because it often relates to pursuit of private commercial interests above wider global benefits (and there are some notable cases of this). However, it is important not to exclude such organisations from GCR governance efforts, although it may be necessary to moderate their influence. They are impacted by global governance, and they can have significant influence on it. If designed well, governance arrangements might motivate companies' contributions to GCR prevention and response.

One model for such action is the UN Global Compact: this invites companies to align their behaviour with international principles and goals in human rights, sustainable development, and social and environmental protection. It currently has participation from over 9,500 companies worldwide. Such a model could be used to raise awareness among companies about GCRs and the behaviours they might adopt to help to address, or at least avoid contributing to such risks.

Companies may engage with global governance on an individual basis or collectively — often through industry organisations. They can play a key role in international standard-setting and in harmonisation and interoperability efforts that extend industry-wide. Such work might help to address some gaps in GCR governance, and as with other transnational actors, they may be able to motivate state action to address particular issues in a timely manner.

The re-insurance industry also has key interests in disaster prevention, resilience, response and recovery, and is another significant actor within GCR governance efforts.

5.3.4 Media organisations

Media organisations are not necessarily deliberate actors in global governance but they can have significant influence on it and a have a key role in GCR governance in terms of communication and public understanding. This role and how it can function constructively during catastrophic events needs to be better understood. Some international organisations provide guidance and/or training on communication during crises (these tend to be aimed at their staff rather than toward media organisations) and have media offices. The UN Office for Disaster Risk Reduction has some relevant initiatives, including a Global Media Network for Disaster Risk Reduction and a Guide for Journalists Covering Disaster Risk Reduction: Disaster through a Different Lens; however, further global guidance developed by and for media organisations around responsible communication and good practice in disaster reporting could have great value. This situation is, of course, complicated by extensive use of social media and continued research efforts in this area are needed, alongside general work to increase public understanding of risk and awareness of misinformation.

There is substantial scope for improving knowledge and understanding about the full range of transnational governance actors and activities that can support GCR governance, building towards recommended actions to enhance and sustain their contributions.

5.3.5 Areas for future research

- More detailed mapping / database of transnational actors across the GCR governance space;
- Case studies of effective practice and areas for shared learning across regimes;
- Legitimacy of transnational actors in global governance;
- Priority which should be given to transnational governance activities within GCR governance;
- Whether there is a relationship between higher levels of transnational actors and activities and effectiveness of governance;
- Whether there is a need to be concerned about areas in which transnational actors dominate GCR governance efforts.

6. Recommendations: Is the International Governance of GCRs Fit for Purpose?

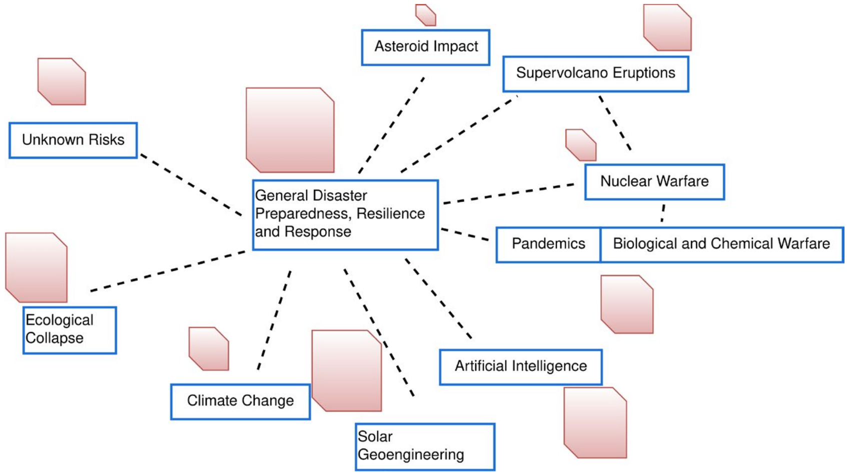
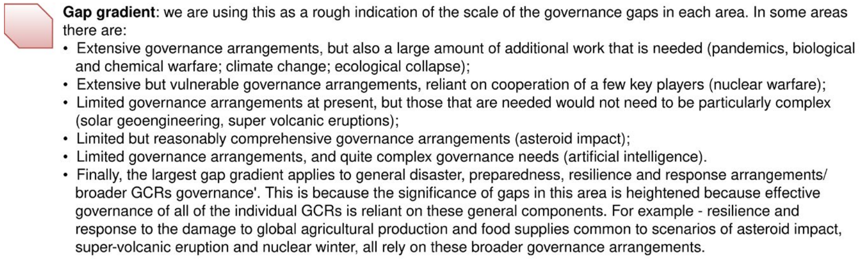


Fig. 13: The gap gradients in global GCR governance.



Gap gradient: we are using this as a rough indication of the scale of the governance gaps in each area. In some areas there are:

- Extensive governance arrangements, but also a large amount of additional work that is needed (pandemics, biological and chemical warfare; climate change; ecological collapse);
- Extensive but vulnerable governance arrangements, reliant on cooperation of a few key players (nuclear warfare);
- Limited governance arrangements at present, but those that are needed would not need to be particularly complex (solar geoengineering, super volcanic eruptions);
- Limited but reasonably comprehensive governance arrangements (asteroid impact);
- Limited governance arrangements, and quite complex governance needs (artificial intelligence).
- Finally, the largest gap gradient applies to general disaster, preparedness, resilience and response arrangements/ broader GCRs governance'. This is because the significance of gaps in this area is heightened because effective governance of all of the individual GCRs is reliant on these general components. For example - resilience and response to the damage to global agricultural production and food supplies common to scenarios of asteroid impact, super-volcanic eruption and nuclear winter, all rely on these broader governance arrangements.

Fig. 14: Gap gradient summary description.

Figure 12 provides an overview of the gaps in different areas of governance (the larger the red icon the more significant and pressing the gap). Figure 13 delivers an overview of the state of the gaps in the different governance areas. This, combined with the summary boxes for each hazard and driver, provides a detailed guide to the strengths and weaknesses of coverage and areas of neglect for each area of GCR governance. This is a high-level overview of the landscape of GCR governance. Each of these areas, particularly larger areas such as ecological collapse and climate change, would require extensive reports of their own to provide a comprehensive analysis.

Something that has not been possible to assess but which may form a vital component of global GCRs governance is the extent to which intelligence agencies cooperate to share information relevant to emerging risks, and whether there are particular actions that might be taken both to improve such coordination and enable some of the information to be shared with other international governance actors.

We suggest the following steps to help advance the state of global GCR governance and fill the gaps:

- Work to identify instruments and policies that can address multiple risks and drivers in tandem;
- Closer research into the relationship between drivers and hazards to create a deeper understanding of our "civilisational boundaries". This should include an understanding of tipping points and zones of uncertainty within each governance problem area;
- Exploration of the potential for "tail risk treaties": agreements that swiftly ramp-up action in the face of early warning signal of catastrophic change (particularly for environmental GCRs);
- Closer examination on the coordination and conflict between different GCR governance areas. If there are areas where acting on one GCR could detrimentally impact another, then a UN-system-wide coordination body could be a useful resource.
- Further work on building the foresight and oversight capacities of the UN for GCRs. More information is needed to investigate whether and on what basis comparison can be made between different areas of GCRs governance in order to prioritise efforts to address gaps. Improving general global preparedness, resilience and response efforts seems an obvious priority because it will contribute to addressing multiple GCRs. However, it is less clear how to prioritise between specific actions that address particular gaps for individual risks.

Notes and References

1 Wilson, Grant. 'Minimizing global catastrophic and existential risks from emerging technologies through international law', *Virginia Environmental Law Journal, 31*(2) (2013): 307–64.

2 Nindler, Reinmar. 'The United Nation's capability to manage existential risks with a focus on Artificial Intelligence', *International Community Law Review, 21*(1) (2019): 5–34. https://doi.org/10.1163/18719732-12341388

3 Pauwels, Eleonore. *The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI* (2019).

4 Bostrom, Nick. 'Existential risks: Analysing human extinction scenarios and related hazards', *Journal of Evolution and Technology, 9* (2002): 1–36.

5 Kemp, Luke. 'Are we on the road to civilization collapse?', *BBC Future* (February 2019).

6 Global Challenges Foundation. *Global Catastrophic Risks Report 2016* (2016).

7 The 2018 GCF Report provides good summary information outlining each of the risk areas, which we do not repeat in full within this report. We would suggest that audiences less familiar with GCRs read the two reports alongside each other.

8 Keohane, Robert O. and David G. Victor. 'The regime complex for climate change', *Perspectives on Politics, 9*(1) (2011): 7–23. https://doi.org/10.1017/S1537592710004068

9 Stone, Peter et al. 'Artificial Intelligence and life in 2030', *One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel* (2016): 1–52. https://ai100.stanford.edu

10 Eilstrup-Sangiovanni, Mette. 'Why the world needs an International Cyberwar Convention', *Philosophy & Technology* (September 2017). https://doi.org/10.1007/s13347-017-0271-5

11 Kunz, Martina and Seán Ó HÉigeartaigh. 'Artificial Intelligence and robotization', in *Oxford Handbook on the International Law of Global Security*, ed. Robin Geiss and Nils Melzer. Oxford University Press (2019), pp. 1–14.

12 Baum, Seth D. 'A survey of Artificial General Intelligence projects for ethics, risk, and policy', *Global Catastrophic Risk Institute Working Paper, 17*(1) (2017): 1–99. https://doi.org/10.2139/ssrn.3070741

13 Whittlestone, Jess et al. *The Role and Limits of Principles in AI Ethics* (2019), pp. 195–200. https://doi.org/10.1145/3306618.3314289

14 Kunz and O hÉigeartaigh (2019).

15 ESA. *Summary of the 13th Meeting of the Space Mission Planning Advisory Group* (*SMPAG*) (2019). https://www.cosmos.esa.int/web/smpag/meeting-13-13-sep-2019-

16 WHO. *Responsible Life Sciences Research for Global Health Security: A Guidance Document* (2010).

17 Pearson, G. S. 'Public perception and risk communication in regard to bioterrorism against animals and plants', *OIE Revue Scientifique et Technique, 25*(1) (2006): 71–82. https://doi.org/10.20506/rst.25.1.1654

18 Crowley, M. *Chemical Control: Regulation of Incapacitating Chemical Agent Weapons, Riot Control Agents and Their Means of Delivery*. Palgrave Macmillan (2016).

19 *UN Security Council Resolutions 1673(2006); 1810 (2008); 1977 (2011); and 2325 (2016)* (n.d.).

20 This guidance is found primarily in: WHO's Manual for the Public Health Management of Chemical Incidents; the WHO Laboratory Biosafety Manual, WHO Laboratory Biosecurity Guidance, and WHO Guidance on Regulations for the Safe Transport of Infectious Substances; and Chapters 5.8 and 6.5 of the Terrestrial Animal Health Code and 1.1.4 of the Manual of Diagnostic Tests and Vaccines for Terrestrial Animals.

21 Knauf, Sascha, Lena Abel and Luisa K. Hallmaier-Wacker. 'The Nagoya Protocol and research on emerging infectious diseases', *Bulletin of the World Health Organization, 97*(6) (2019): 379. https://doi.org/10.2471/BLT.19.232173; WHO, *Public Health Implications of Implementation of the Nagoya Protocol* (2019). https://www.who.int/activities/public-health-implications-of-implementation-of-the-nagoya-protocol

22 Nuclear Threat Initiative. *Global Health Security Index* (2019). https://www.ghsindex.org./

23 Kemp, Luke. 'Better out than in', *Nature Climate Change, 7* (2017): 458–60. https://doi.org/10.1038/nclimate3309

24 Kemp, Luke. 'US-proofing the Paris Climate Agreement', *Climate Policy, 17*(1) (2017): 86–101. https://doi.org/10.1080/14693062.2016.1176007

25 GCF. *Green Climate Fund* (2019). https://www.greenclimate.fund/home

26 Steffen, Will et al. 'Trajectories of the Earth system in the Anthropocene', *Proceedings of the National Academy of Sciences, 115*(33) (2018): 8252–59. https://doi.org/10.1073/pnas.1810141115

27 Lenton, Timothy M. 'Early warning of climate tipping points', *Nature Climate Change, 1* (2011): 201–9. https://doi.org/10.1038/nclimate1143

28 Green, Fergus and Richard Denniss. 'Cutting with both arms of the scissors: the economic and political case for restrictive supply-side climate policies', *Climatic Change* (2018). https://doi.org/10.1007/s10584-018-2162-x; Green, Fergus. 'Anti-fossil fuel norms', *Climatic Change, 150*(1–2) (2018): 103–16. https://doi.org/10.1007/s10584-017-2134-6

29 Newell, Peter and Andrew Simms. 'Towards a fossil fuel non-proliferation treaty', *Climate Policy* (2019): 1–12. https://doi.org/10.1080/14693062.2019.1636759

30 Rafferty, A. et al. 'Less than 2 °c warming by 2100 unlikely', *Nature Climate Change, 7* (2017): 637–41. https://doi.org/10.1002/cncr.27633.Percutaneous

31 CAT. '2100 warming projections', *Climate Action Tracker* (2019). https://climateactiontracker.org/global/temperatures/

32 Rogelj, Joeri et al. 'Perspective : Paris Agreement climate proposals need boost to keep warming well below 2 ° C', *Nature Climate Change, 534* (June 2016): 631–39. https://doi.org/10.1038/nature18307

33 Kemp, Luke. *A Systems Critique of the 2015 Paris Agreement on Climate, Pathways to a Sustainable Economy: Bridging the Gap Between Paris Climate Change Commitments and Net Zero Emissions* (2017). https://doi.org/10.1007/978-3-319-67702-6_3

34 Talberg, Anita et al. 'Geoengineering governance-by-default: An Earth system governance perspective', *International Environmental Agreements: Politics, Law and Economics, 18*(2) (2018): 229–53. https://doi.org/10.1007/s10784-017-9374-9

35 Pasztor, Janos et al. *Geoengineering: The Need for Governance* (2019); Talberg et al. (2018); Brent, Kerryn, Jeffrey McGee and Jan McDonald. 'The governance of geoengineering: An emerging challenge for international and domestic legal systems?', *Journal of Law, Information and Science, 24*(1) (2015): 1–33.

36 ICJ Rep 221. *Legality of the Threat or Use of Nuclear Weapons* (*Advisory Opinion*) (1996).

37 Brent, McGee and McDonald (2015).

38 Fleming, James Rodger. 'The pathological history of weather and climate modification: three cycles of promise and hype', *Historical Studies in the Physical and Biological Sciences, 37*(1) (2006): 3–25. https://doi.org/10.1525/hsps.2006.37.1.3

39 Convention on Biological Diversity. *Climate-Related Geoengineering and Biodiversity* (2019). https://www.cbd.int/climate/geoengineering/

40 Pasztor et al. (2019).

41 Brent, McGee and McDonald (2015).

42 Currie, Adrian. 'Geoengineering tensions', *Futures, 102* (2018): 78–88. https://doi.org/10.1016/j.futures.2018.02.002

43 Avin, Shahar et al. 'Classifying Global Catastrophic Risks', *Futures* (2018). https://doi.org/10.1016/j.futures.2018.02.001

44 UN CEB. *Report of the High-Level Committee on Programmes at Its Thirty-Seventh Session* (2019).

45 GCF. *Global Catastrophic Risks Report 2018* (2018).

46 Kemp (2019).

47 Scheidel, Walter. *The Great Leveler: Violence and the History of Inequality From the Stone Age to the Twenty-First Century*. Princeton University Press (2017).

48 Turchin, Peter. *War and Peace and War: The Rise and Fall of Empires*. Plume (2016).

49 Hauner, Thomas, Branko Milanovich and Suresh Naiduz. *Inequality, Foreign Investment, and Imperialism* (2017).

50 Hickel, Jason. *The Divide: Global Inequality from Conquest to Free Markets*. W. W. Norton & Company (2018).

51 UN. *Transforming Our World: The 2030 Agenda for Sustainable Development* (2015).

52 UN Inter-Agency Task Force on Financing for Development. *International Efforts to Combat Tax Avoidance and Evasion* (2019). https://developmentfinance.un.org/international-efforts-combat-tax-avoidance-and-evasion

53 Piketty, Thomas. *Capital in the Twenty-First Century*. Belknap Press (2017).

54 Authors calculations using World Bank data.

55 Verbeek, Jos and Israel Osorio Rodarte. *Increasingly, Inequality Within, Not Across, Countries Is Rising* (2015); Scheidel (2017).

56 Kemp (2019).

57 Zucman, Gabriel. 'Global wealth inequality', *National Bureau of Economic Research Working Paper Series, 25462* (2019): 1–45.

58 Rockström, Johan et al. 'Planetary boundaries: Exploring the safe operating space for humanity', *Ecology and Society, 461*(24) (2009): 472–75. https://doi.org/10.5751/ES-03180-140232

59 IEA Database Project. *International Environmental Agreements (IEA) Database Project* (2019). https://iea.uoregon.edu/

60 GEF. 'Funding', *Global Environment Facility* (2019). http://www.thegef.org/about/funding

61 GEF.

62 Morseletto, Piero. 'Confronting the nitrogen challenge: Options for governance and target setting', *Global Environmental Change, 54* (2019): 40–49. https://doi.org/10.1016/j.gloenvcha.2018.10.010

63 Morseletto (2019).

64 Ecological Footprint Network. *Ecological Footprint* (2019). https://www.footprintnetwork.org/our-work/ecological-footprint/

65 WWF. *Living Planet Report — 2018: Aiming Higher*, ed. M. Grooten and R.E.A. Almond (2018).

66 Bastide, L. 'Crisis communication during the Ebola outbreak in West Africa: The paradoxes of decontextualized contextualization', in *Risks Communication for the Future*, ed. M. Bourrier and C. Bieder. Springer (2018): 1–175.

67 Huairou Commission. *Huairou Connects Grassroots Women to Support Community-Led Development* (2019). https://huairou.org/network/community-practitioners-platform/

68 Go Fossil Free. *Divestment Commitments* (2018). https://gofossilfree.org/divestment/commitments/

69 Mayors for Peace. *About Us* (2019). http://www.mayorsforpeace.org/english/outlines/index.html

20. The Stepping Stones Approach to Nuclear Disarmament Diplomacy[1]

Paul Ingram

Highlights:

- This chapter provides a personal and reflective account of the author's efforts in the field of nuclear disarmament diplomacy. It focuses on the development of the Stepping Stones Approach.
- The Stepping Stones Approach is an iterative approach that starts from the point of radical visions for the future, striving towards common security and greater collaboration. Through dialogue and the iterative development of proposals and agreements, the approach builds towards incremental action.
- The Stepping Stones Approach is an effort to transform diplomacy away from zero-sum confrontational and positional negotiation, towards more adaptable and exploratory engagements. It uses a form of incrementalism to develop ambitious proposals for change. It has emerged because power in the international system in relation to nuclear diplomacy is highly concentrated in the hands of the nuclear armed states.
- The approach may well carry lessons for other fields of catastrophic risk — where extant modes of political or institutional engagement are frustrated by power competition, political disagreement or seemingly irreconcilable priorities.

https://doi.org/10.11647/OBP.0360.20

This chapter was specially written for this volume and sets out a roadmap for how change can be enacted in contested global diplomatic efforts to respond to extreme global risk. A further guide to engaging with policy-makers and stakeholders is contained in Chapter 18.

Introduction

The Stepping Stones Approach (the Approach) was developed by the author and first adopted by a coalition of sixteen governments in June 2019 in order to break the deadlock in nuclear disarmament diplomacy.

The Approach arises out of frustration and some desperation. The ambition for complete nuclear disarmament is articulated in Article VI of the nuclear Non-proliferation Treaty (NPT), as well as the more recent Treaty for the Prohibition of Nuclear Weapons.[2] Yet all nine nuclear armed states are modernising their nuclear weapons. Having exercised significant strategic restraint, China is now expanding its arsenal.[3] There is a well-established US view that strategic competition and war with China may well be inevitable.[4]

The international community last agreed to an integrated nuclear disarmament strategy at the 2010 NPT Review Conference.[5] This opened the door to three international conferences on humanitarian consequences of nuclear weapons in 2013–14 and indirectly to negotiations on the Treaty for the Prohibition of Nuclear Weapons (TPNW) adopted on 7th July 2017.[6] As of November 2022 this treaty had 91 signatories and 68 state parties, but no nuclear armed states or any state in alliance with one.[7] This may have turned up the heat but progress on achieving actual disarmament has been absent.

Most efforts to drive disarmament diplomacy tend to focus on the dangers, ethics and legal obligations arising from previous agreements, but fail to account sufficiently for the attachment to the security and influence that nuclear weapons are perceived to convey. These in turn are built upon assumptions that bear some scrutiny. These include that possession and threat delivers effective and unique deterrence, that the risk is acceptable, that great powers have responsibilities to control global outcomes, and that strategic competitors will contemplate extreme measures for advantage. Successful moves to drive disarmament require states to address these competing assumptions and commitments, and to find alternative less dangerous means to achieve their objectives.[8]

When entering international nuclear weapon negotiations, officials prepare for a confrontational experience as they weigh up their opposition and the competing interests involved. More often than not they carry a scepticism around the prospects for progress. The Approach seeks to change this negotiating culture. It involves officials seeking out opportunities to explore possible futures as a means to collect ideas for early interventions, even as they accept the complexities that resist solutions. These ideas are then used to prompt open dialogue between key stakeholders with a view to attempting to settle on early, modest action.

Taking a systems view of change, the Approach grew out of a desire to escape polarised debate and build concrete improvements in nuclear disarmament diplomacy. The Approach arises out of an awareness that states co-exist in interdependent common security relationships, and that efforts to improve relationship and reduce fear benefit security. It thus draws states into *a process* that *moves towards* an alternative paradigm of common security.[9]

This chapter begins by explaining the method of the Stepping Stones Approach. It then outlines the essential ingredients of the associated culture, based upon an appreciation of:

- systems, emergence and complexity;
- polarity management in which binaries that drive conflict come to be seen as framings that can brings people together; and
- relationship and process being critical to outcomes.

1. The Method

The Approach involves a number of elements or steps in a non-linear iterative process (illustrated in Figure 1):

1. **Radical visions for the future**, striving towards common security and greater collaboration.
2. **Analysis, acceptance, and pluralism**, understanding the complexities using an inclusive dialogue.

3. **Proposals and dialogue**, view to triggering further iterations of proposals and ideas.
4. **Early, modest action** that is incremental, usually taken by nuclear weapon states.
5. **Review, evaluate and adapt**, as transformation emerges in a non-linear and unpredictable manner.

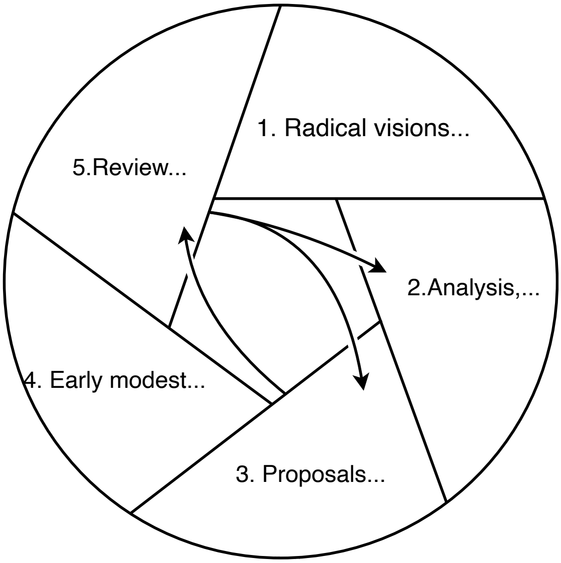


Fig. 1: Diagram of the Approach.

1.1 Visions for the future

The Approach starts much like any strategy, contemplating the desired qualities of the world we seek to inhabit. But this is not a settling on a particular outcome. Effective and sustainable action comes from inspiration around potential futures rather than a rejection of the present. The Approach encourages those looking for change to develop and communicate constructive visions for how things could be, but to hold them lightly. If we attach too strongly to a particular vision we are very likely to seek the means to drive the system in our direction, to be inflexible and to drive conflict with others who do not share it or our perspective. These visions are not manifestos, commitments, or targets, but rather guide-stars — ideas that help develop and communicate a desirable direction of travel, elements of progress and thee values that underpin them. They will adapt as the international context evolves, as we discover new directions that could better meet our collective objectives. An important challenge when we engage in such visioning is to retain that sense of adaptability.

Talking about the visions helps us better communicate with each other about our shared purpose and values and assists in the exploration of initial steps in those directions. By distinguishing those visions from the more modest immediate policy actions that are to be implemented in the immediate term, we can draw the sting from some of the conflicts that stymie progress and give explicit encouragement to those who advocate radical visions.

When we advocate for common values such as equity, justice, human rights, and responsive governance we need to do so from a place of openness and respect, understanding and owning our own failings and drawing the other into dialogue. This search for common ground in fundamental values lies at the heart of international society. Genuine and admirable attempts have been made in the last century to articulate and develop such values, formalised after World War II in the Charter of the United Nations and the Universal Declaration of Human Rights, and more recently in the UN's Sustainable Development Goals for 2030.[10] Recognition of the power of shared values is a critical lubricant in the machinery of international dialogue that is weakened most by cynicism or the view that there only exists self-interest.

1.2 Analysis and acceptance of the situation, adopting pluralism

The Approach has at its core an appreciation of the contribution diversity has to sustainable change when engaging stakeholders, an acceptance of the complexities involved and a resistance to the ubiquitous temptation to over-simplify and rush to judgement. [11] Decisions are stronger, richer, and more sustainable when a variety of perspectives are engaged.

Seeking out and engaging with diverse views is an antidote to the righteous group-think tendencies and confirmation bias that so often harms genuine dialogue and effective policy creation.[12] No one person or state has a monopoly on the truth, which is dynamic, multi-faceted, with tensions, polarities, and contradictions.

If nuclear disarmament is to come from diplomacy and improvements to global security, it will require constructive and voluntary steps by the nuclear weapon states, taken in good faith and with confidence. Disarmament proposals made in this context must be considered in relation to existing nuclear deterrence postures.

Pursuit of a belief in disarmament or deterrence in a dogged and inflexible manner will often trigger resistance. Many people believe disarmament is the most effective, long-term solution to improving global security. Indeed, reducing the role of nuclear weapons in national security strategies is central to the Stockholm Initiative, and necessary for progress in nuclear disarmament. This is based upon the belief that moves to reduce reliance on nuclear deterrence can send positive signals of intent to improve strategic relationships and reduce nuclear risks.

This agenda also appears to have been the most challenging part of the Initiative's agenda for some of the nuclear weapon states in private consultations, even as they have supported the Initiative as a whole and have similar objectives in their national nuclear postures.[13] Some within the nuclear weapon states believe that reductions in nuclear salience at a time of strategic competition can damage nuclear deterrence because they could be interpreted as weakening resolve for nuclear use, and thus could perversely increase nuclear risks by emboldening aggression. They believe that reduced nuclear salience needs to *follow* improved strategic relationships between nuclear armed states, rather than seek to improve relationships by reducing nuclear salience within strategic defence postures first. This is reflected in their focus upon what they describe as creating a strategic environment conducive to nuclear disarmament.[14]

The disagreement between these two positions should not be understated, but there is scope for progress when we acknowledge the strengths and weaknesses of both. This will not resolve the contradiction but will encourage those participating to pay attention to others, to the evidence, and a wider variation of possibility. After all, the objective of *security* is shared

Whilst outright confrontation is usually counterproductive, the Approach does involve drawing attention to some of the nuclear weapon states most dangerous and escalatory behaviours and encouraging them to engage in open and respectful discussion about these actions as a first step on the road towards disarmament. This might include, for example, the policy of launch-under-attack, which many believe presents the greatest risk of nuclear exchange arising from a false alarm on the basis that credible warnings of incoming long-range missiles create a use-them-or-lose-them situation.[15]

1.3 Proposals and dialogue

Trust and confidence take time to build up, and state representatives need to feel their nation's concerns and priorities are heard and respected in the process, enabling them to witness the mutual benefits that can arise before they are willing to invest further in shared governance. Proposals therefore need to account for the interests and perspectives of all main stakeholders with a view to drawing them into dialogue, an open process that involves joint exploration of the landscape and possible improvements that could be attempted. Proposals are best tabled as invitations to explore and participate. If a state responds with a counter-proposal, that itself is a recognition of the process and a success. Engaging constructively with such a counter-proposal and seeking to integrate core objectives is the stuff of successful diplomacy.

When considering steps to progress disarmament we need to understand the drivers behind the dysfunctional relationships that underpin nuclear deterrence rather than continue attempting to keep the lid on the situation.[16] We do well when we draw states that challenge us into a process that involves patient attempts to develop mutual respect and genuine attempts to break the cycles of violence.

1.4 Practical, incremental action

The Approach involves taking *early* practical steps with the intention of building momentum, understanding that there will be dead ends and false starts. In complex environments, even very small steps can have unpredictable impacts upon other parts of the landscape that can then open up new challenges and opportunities. The four statesmen that reignited interest in global nuclear disarmament amongst the elites within nuclear weapon states, Schultz, Perry, Kissinger and Nunn, used the analogy of climbing a mountain in their seminal letter to the Wall Street Journal in January 2007.[17] When ascending the slopes, a climber sees new opportunities and challenges as they gain height. Unfortunately, we seem to have fallen down several crevasses in the last decade. As a result, confidence in the step-by-step approach has been damaged.

We could see ourselves as existing in a metaphorical landscape of peaks and valleys, and efforts to shift to another dynamic equilibrium in a new valley takes larger nudges, or a series of small ones, because of the negative feedback loops that return our systems to the *status quo* and keep them stable. We might imagine a multitude of potential equilibrium points within international strategic relations.[18] In terms of the Approach, we need to consider individual stepping stones in their own right, but it will take implementing a series of them, likely in a number of areas, to unlock stickiness in the system and achieve sustained progress towards nuclear disarmament. We have to expect resistance, failures that take us back to an equilibrium we were hoping to shift. But when we build momentum we have a hope of driving lasting change.

1.5 Adapt as events unfold, engage with emergence

The Approach involves improvements, not solutions. As we achieve them, further possible improvements will emerge. It is a feature of the process that transformation emerges in a non-linear and unpredictable manner — emergent change.[19]

The Approach is more likely to be successful if all parties see the dialogue as a shared learning process that involves concrete implementation by the nuclear weapon states. The diagram below is a representation of a learning cycle involved in nuclear diplomacy and implementation, reflecting the fact that the changes themselves (to the right half of Figure 2, in blue) will happen nationally within those states that possess nuclear weapons.

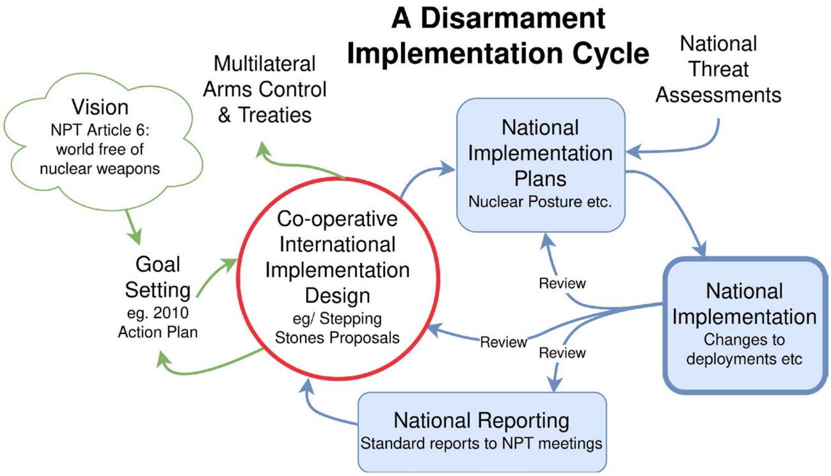

A Disarmament Implementation Cycle
Vision
NPT Article 6: world free of nuclear weapons
Multilateral Arms Control & Treaties
National Threat Assessments
National Implementation Plans
Nuclear Posture etc.
Goal Setting
eg. 2010 Action Plan
Co-operative International Implementation Design
eg/ Stepping Stones Proposals
Review
Review
Review
National Implementation
Changes to deployments etc
National Reporting
Standard reports to NPT meetings

Fig. 2: The learning cycle involved in nuclear diplomacy and implementation.

2. Essentials Behind the Approach

2.1 It takes appreciation of complexity and emergence

Complex and chaotic systems display emergent properties. Their wicked problems demand, "an approach that requires experimentation and the capacity to allow a path forward to emerge over time — the common cause-and-effect thinking and tools that leaders use to fix problems don't create the results they expect".[20] Objectives appear contradictory, are impossible to fully comprehend, and defy efforts to simplify or to control.[21]

Signals and reactions can be unpredictable. For example, significant disarmament moves can signal a confidence in the future and reduced threat, encouraging competitors to relax and de-escalate. On the other hand, they can be interpreted as demonstrating less resolve or even weakness, encouraging an assertive competitor to move into the space created. Clarity and consistency of messaging reduces the possibility of misinterpretation but does not eradicate it. It helps to ensure our analysis and actions considers the broader context, the imperfections, unintended consequences, our political and cognitive distortions, and the likely systems failures. It requires a humility rarely displayed by leaderships and underappreciated by their publics. The complexities and uncertainties are a powerful reason to proceed with sensitivity to the feedback signals, but also with confidence that making interventions enables learning and growth within the system.

The Approach is not a conventional strategy, planning and then implementing steps. The vision and the steps emerge and adapt as we interact. This requires of us that we recognise that we, individually and collectively, are not in control of outcomes, and that we are engaged in a learning cycle. It means that we have to be more flexible with our personal or political attachments to particular outcomes.

When engaging with complex, emergent systems the attempt to grip and take control can exacerbate the conflict. People often think, for example, that the most effective negotiating strategy is first to build up one's own bargaining position by accumulating assets or positions that can later be traded, or through sheer force of argument, and developing a hard reputation for inflexibility. Sometimes this strategy can drag out the start of serious negotiations for many years as antagonists square off against one another, issue threats and impose penalties.[22] In contrast, the adaptive and collaborative exploration of possibilities opens up unforeseen possibilities and increases the chance of serving the common interest.

2.2 Working across polarities as a strategy

The Approach is designed to steer global collaboration on the nuclear disarmament agenda, for which there is an agreed but stalled programme of action.[23] It seeks to answer the "how" in terms of effective diplomacy and engagement. The "what", its agenda, is co-created in the process. [24]

People often use rational argument, manipulate evidence, incentives and emotion, threaten or punish to build support for their preferred solutions. This can descend into a simple trial of strength. We externalise blame, express our anger, and feel the righteousness of our beliefs.[25] We often believe conflict is necessary to resolve disputes, and we institutionalise it across many arenas in life (such as democratic debate, in the courts, or generally asserting our interests). This can drop into a perpetual cycle of conflict between entrenched positions. These approaches to change are particularly ineffective when pursued from a position of weakness, as those with greater power usually use that power to protect the *status quo*.

But even they can attract fierce resistance when they attempt to exert their will, such that any benefit is degraded or eliminated.

We often observe the tendency for complex situations to exhibit two or more polarities in tension with one another, each with advantages and disadvantages for people within the community, but that do not exist in isolation.[26] Too often people pick a solution from one or other polarity, when the challenge is about managing the dynamic balance. When people and states surface and acknowledge the tensions between the polarities, it can strengthen dialogue and understanding without requiring people to switch their position or even to compromise. We see our perspective within a broader context and deepens an appreciation of the dynamic nature of the system as it changes over time. One such example of polarity is transparency and ambiguity.

2.2.1 Transparency and ambiguity

Transparency over nuclear arsenals, doctrine and intentions impacts international stability and builds trust, but sits uncomfortably with ambiguity in nuclear deterrence. Clarity and active management of strategic relationships can reduce nuclear risk arising from misunderstanding or misperception. Alongside inspections and verification, it is an essential ingredient of multilateral nuclear diplomacy and arms control. Transparency shows respect for others, brings trust and stability, and enables others to become friendly critics rather than hostile challengers. It reduces the risks of strategic surprise, builds confidence, and thereby facilitates lower defence spending. Greater transparency also facilitates communication of genuine intent and resolve when necessary.

On the negative side, transparency over nuclear use may give comfort to aggressors if they believe that smaller transgressions would go unpunished, or force one's own hand when they break red lines. Being open about deployments can expose them to action that neutralises their impact, or expose one's own weaknesses and vulnerabilities. Reversal of transparency in crisis might further escalate tensions at sensitive moments.

The perceived benefits of secrecy and ambiguity are so strong that they are the default behaviour of many governments.[27] Military leaders often oppose formal statements that limit options before a crisis. Some think them naïve as they vanish under existential crisis.[28] Ambiguity delivers doubt in the minds of an aggressor and complicates their strategic planning. Opacity might deliver some additional deterrence against other unspecified threats, or broader influence over international outcomes. Ambiguity can come at considerable additional cost. A government insisting all options remain on the table, for example, invites strong push back and damages international law when it implies a willingness to operate beyond it. Ambiguity can undermine confidence of allies, undermine global diplomacy, and can suggest an indefinite and inflexible attachment to nuclear deterrence.

These features are summarised in the table below:

Table 1: An overview of the positives and negatives of transparency and ambiguity in nuclear diplomacy.

TRANSPARENCY	AMBIGUITY
Positives	**Positives**
Provides trust, clarity and confidence Facilitates collective understanding & management of strategic relationships Clearer signalling Essential ingredient of nuclear diplomacy and arms control Shows respect for international community	Freedom of action in crisis Maximises return from investment Doubt in the minds of aggressors, complicating their calculations Delivers additional deterrence/ influence
Negatives	**Negatives**
Gives comfort to aggressors operating under the red line Force own hand when aggression above the line Gives away valuable strategic info Reversals in transparency in crises could escalate the conflict Lack of credibility in making peacetime promises	Higher risk of misunderstanding/ misperception Invites strong responses Damages international law/ cooperation Undermines allies' confidence Is an obstacle to disarmament and signals an apparent indefinite commitment to deterrence

Understanding and exploring in good faith the positives and negatives of the two polarities of transparency and ambiguity enables a more nuanced discussion of the options, and a recognition that this is more about managing objectives in tension rather than the ideological struggles these debates are often characterised as.

2.3 Attention to process and relationship

In addition, the Approach involves diplomats paying as much attention to the *process* of disarmament diplomacy as to its *content*, establishing and deepening the interpersonal relationships with officials from other states, particularly those with competing interests or different perspectives.[29] Paying attention to and valuing those interconnections, building trust, confidence and understanding, even in the face of challenging dynamics and conflicting interests and ideologies, are critical steps in finding breakthroughs.

Whilst those using the Approach individually or collectively may have particularly high hopes and radical ambition, it involves a pragmatic, collective, respectful and adaptive learning cycle. The Approach seeks to draw the nuclear weapon states into a progression of steps they can consent to. It respects and values everyone's perspective recognising that when people engage fully, many of the underlying objectives are already met. It encourages states to hold their positions less tightly and see the broader context within which they and their neighbours co-exist.

Conclusion

The Stepping Stones Approach is an effort to transform diplomacy away from zero-sum confrontational and positional negotiation, towards more adaptable and exploratory engagements. It uses a form of incrementalism to develop ambitious proposals for change. It has emerged because power in the international system in relation to nuclear diplomacy is highly concentrated in the hands of the nuclear armed states, who have not delivered the level of progress envisaged in the 2010 Action Plan. There is only so far that appeals delivered as speeches in intergovernmental conferences can go before people start looking for other approaches. Yet, when power is concentrated, it does not generally pay for the less powerful to force the issue, but rather to draw those with power into a process whereby all come to recognise the improvements to the system as a whole. Progress on nuclear disarmament is an imperative to our collective survival, and yet is not happening. It may become increasingly critical as other stresses on systems of global governance rise. The evolution of those systems, and the cultures that support them, needs to speed up significantly, and the Stepping Stones Approach is just one attempt to do so.

Notes and References

[1] This chapter is drawn extensively from a report: Ingram, Paul. *The Stepping Stones Approach to Nuclear Disarmament Diplomacy: A Personal Explanation of the Approach From One of Its Designers*. British American Security Information Council (BASIC) (December 2021).

[2] The text of the Non-Proliferation Treaty is available on the United Nations Office of Disarmament Affairs website. 'Treaty on the Non-Proliferation of Nuclear Weapons'. https://www.un.org/disarmament/wmd/nuclear/npt/text/

[3] US intelligence is projecting Chinese modernisation of its nuclear forces and an increase from 400 at end 2022 to 1500 warheads by 2035. US Department of Defense. *China Military Power Report 2022* (2022), p.94. https://www.defense.gov/CMPR/

[4] Allison, Graham. *Destined for War: Can America and China Escape Thucydides's Trap*? (2017).

[5] *Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons Final Document, Volume I* (2010), pp. 19–24. https://www.un.org/en/conf/npt/2010/

[6] The final document included for the first time a reference to "the catastrophic humanitarian consequences that would result from the use of nuclear weapons". *Ibid*., p. 19, paragraph 80. The vote to adopt the TPNW was 122 to 1 with 1 abstention, though the nuclear armed states and almost all their allies failed to participate. 'United Nations Conference to negotiate a legally binding instrument to prohibit nuclear weapons, leading towards their total elimination', *United Nations*. https://www.un.org/disarmament/tpnw/

[7] Nuclear armed states and allies deny the TPNW has any relevance to them. See, for example, statement by Benjamin Sharoni, Israel Ministry of Foreign Affairs to the First Committee of the UN General Assembly, on October 31st: https://reachingcriticalwill.org/images/documents/Disarmament-fora/1com/1com22/eov/L17_Israel.pdf

[8] All well-established systems have mechanisms that act as negative feedback loops or checks on rapid change. These drive the system back towards its natural equilibrium, even when change is attempted by those in positions of strong formal power. Without such negative feedback features, the system would have been unstable and would previously have spun off into another state. Harvard professors Lisa Laskow Lahey and Robert Kegan talk of "immunities to change" in our lives, as individuals, groups or nations. Kegan, Robert and Lisa Laskow Lahey. *Immunity to Change: How to Overcome It and Unlock the Potential in Yourself and Your Organization*. Harvard Business Review Press (2009).

[9] 'Common Security 2022: For our Shared Future', https://www.ituc-csi.org/IMG/pdf/commonsecurity_report_2022_final.pdf

[10] For links to the UN Charter, the Declaration of Human Rights and the Sustainable Development Goals: the text of the Non-Proliferation Treaty is https://www.un.org/en/about-us/un-charter; https://www.un.org/en/about-us/universal-declaration-of-human-rights and https://sdgs.un.org/goals

[11] This was an observation made by Chris Ford in his challenging article, 'The Bodhisattva Vow and Nuclear Arms', *Upaya Newsletter* (July 2009). He wrote this some years before joining President Trump's White House and then becoming US Under-Secretary of State.

[12] Ingram, Paul. 'Addressing our worst global nightmares whilst managing our righteousness', *The Friends Quarterly, 55*(4) (November 2022), p.12.

[13] It is also an explicit top-level objective of the 2022 US Nuclear Posture Review (page 1).

[14] Under-Secretary of State Chris Ford first proposed an initiative to create the conditions for nuclear disarmament in March 2018, and the first inaugural meeting of CEND was held in July 2019. See Potter, William. *Taking the Pulse at the Inaugural Meeting of the CEND Initiative* (July 2019). https://nonproliferation.org/taking-the-pulse-at-the-inaugural-meeting-of-the-cend-initiative

[15] Perry, William J. and Tom Z. Collina. *The Button, the New Nuclear Arms Race and Presidential Power From Truman to Trump*. BenBella Books (2020).

[16] Rogers, Paul. *Losing Control: Global Security in the 21st Century*. Pluto (2010).

[17] Schultz, George P., William J. Perry, Henry A. Kissinger and Sam Nunn. 'A world free of nuclear weapons', *Wall Street Journal* (4 January 2007).

[18] Young, Ed. *I Contain Multitudes: The Microbes Within Us and a Grander View of Life*. Harper Collins (2016), p.117. Emphasis added.

[19] Brown, Adrienne Maree. *Emergent Strategy: Shaping Change, Changing Worlds*. AK Press (2017).

[20] Clark, Larry. *Navigating Complexity: A New Map for a New Territory*. Harvard Business Publishing (November 2018). https://www.harvardbusiness.org/navigating-complexity-a-new-map-for-a-new-territory/

[21] Emergent properties are those that are not inherent within the components of a system but that are observed when the system as a whole operates. For example, human cells each have functions, and collectively make up organs and other parts of the body that all together make up an individual.

[22] This is well illustrated in the stand-off between the United States and Iran over its nuclear enrichment programme, particularly from 2005 to the secret negotiations in Oman in 2013. Countless efforts to find middle ground compromises that met the core needs of both parties were made, including an influential track two series facilitated by the author and colleagues, but the two sides were intent on building up their capacities (Iran its enrichment programme, the United States their sanctions regime) in order to later come to the negotiating table with strong hands.

[23] This commitment is expressed both in the NPT Treaty itself as well as in subsequent Review Conference agreements and consensus documents.

[24] The "what" in the case of the Stockholm Initiative is contained within the Sweden NPT working paper referred to above, and in the Annex to the Ministerial Declaration

issued at the Berlin meeting in March 2020: *Stepping Stones for Advancing Nuclear Disarmament – Annex to the Declaration of the Berlin Ministerial Meeting* (25 February 2020). https://www.government.se/497342/globalassets/regeringen/lena-micko-test/stepping-stones-for-advancing-nuclear-disarmament.pdf

[25] Haidt, Jonathan. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon (2012); Ingram (2022), p. 12.

[26] Johnson, Barry. *Polarity Management: Identifying and Managing Unsolvable Problems* (2nd edition). HRD Press (2014). Polarities Management was introduced to BASIC and Emergent Change by the Nucleus Group (www.thenucleusgroup.com), who helped us develop a series of workshops with diplomats and officials from a number of countries under a grant from the Carnegie Corporation of New York in 2019.

[27] The idea of ambiguity (that applies to the possibility of nuclear use) can sometimes be used to mask secrecy (that surrounds capabilities and their deployments). When publishing its Integrated Review in March 2021 the UK government abandoned its practice of operational transparency (in regards to the numbers of warheads and missiles on patrol) and stepped into greater secrecy. It misleadingly described the move as an extension of its "long-standing policy of deliberate ambiguity... [that] complicates the calculation of potential aggressors, reduces the risk of deliberate nuclear use by those seeking a first-strike advantage, and contributes to strategic stability". See *Global Britain in a Competitive Age: The Integrated Review of Security, Defence, Development and Foreign Policy* (Cabinet Office Policy Paper). HMG (March 2021), p. 77.

[28] Roberts, Brad. 'Debating nuclear no-first-use, again', *Survival, 61*(3): 39–56. https://doi.org/10.1080/00396338.2019.1614788

[29] Lederach, John Paul. *The Moral Imagination: The Art and Soul of Building Peace*. Oxford University Press (2010).

21. It Takes a Village: The Shared Responsibility of “Raising” an Autonomous Weapon

Amritha Jayanti and Shahar Avin

Highlights:

- Expectations around future capabilities of Lethal Autonomous Weapons Systems (LAWS) have raised concerns for military risks, ethics, and accountability. The UK’s position has attempted to address these concerns through a focused look at the weapons review process, human-machine teaming (or ”meaningful human control”), and the ability of autonomous systems to adhere to the Rules of Engagement.
- Further, the UK has stated that the existing governance structures — both domestic and international — around weapons systems are sufficient to deal with concerns around the development, deployment, and accountability for emerging LAWS, with no need for novel agreements on the control of these weapons systems.
- In an effort to better understand and test the UK’s position on LAWS, the CSER ran a research project that interviewed experts in multiple relevant organisations, structured around a mock parliamentary inquiry of a hypothetical LAWS-related civilian death.

https://doi.org/10.11647/OBP.0360.21

- The responses highlighted different conceptions of future systems, which were sometimes complementary but sometimes contradictory, as well as challenges and accountability measures. They have provided rich "on the ground" perspectives and highlight the very wide range of intervention points where humans are expected, and should be supported, to make decisions that enable legal, safe, and ethical weapon systems. These all need to be considered by any military that is considering acquisition and deployment of autonomous and semi-autonomous weapon systems.

This chapter was initially presented as a workshop paper in 2020. Using expert interviews and scenarios, the chapter provides an empirically informed account of the multiple points at which meaningful human oversight and control of autonomous weapons ought to be exercised. Similar methodological approaches are presented in several chapters of this volume, including 8, 16 and 14.

1. Introduction

With the increasing integration of digital capabilities in military technologies, many spheres of the public — from academics to policy-makers to legal experts to nonprofit organisations — have voiced concerns about the governance of more "autonomous" weapons systems. The question of whether autonomous weapons systems pose novel risks to the integrity of governance, especially as it depends so heavily on the concept of human control, responsibility, and accountability, has become central to the conversations.

The United Kingdom (UK) has posited that lethal autonomous weapons (LAWS), in their current and foreseeable form, do not introduce weaknesses in governance; existing governance and accountability systems are sufficient to manage the research, development, and deployment of such systems and the most important thing we can do is focus on improving our human-machine teaming. Our research project seeks to test this theory by asking: with the introduction of increasingly autonomous agents in war (lethal autonomous weapons/LAWS), are the current governance structures (legal, organisational, social) in fact sufficient for retaining appropriate governance and accountability in the UK MoD? By attempting to confront strengths and weaknesses of existing governance systems as they apply to LAWS through a mock parliamentary inquiry, the project uncovers opportunities for governance improvements within Western military systems, such as the UK.

2. Background

Computers and algorithms are playing a larger and larger role in modern warfare. Starting around 2007 with writings by Noel Sharkey, a roboticist who heavily discusses the reality of robot war, members of the research community have argued that the transition in military technology research, development, and acquisition to more autonomous systems has significant, yet largely ignored, moral implications for how effectively states can implement the laws of war.[1] Segments of this community are concerned with the ethics of decision-making by autonomous systems, while other segments believe the key concern is regarding accountability: how responsibility for mistakes is to be allocated and punished. Other concerns raised in this context, e.g. the effects of autonomous weapon systems on the likelihood of war, proliferation to non-state actors, and strategic stability, are beyond the scope of this brief, though they also merit attention.

2.1 UK position on LAWS

The United Kingdom's representatives at the UN Group of Governmental Experts (GGE) on Lethal Autonomous Weapon Systems (LAWS) have stated that the UK believes the discussions should "continue to focus on the need for human control over weapon systems and that the GGE should seek agreement on what elements of control over weapon systems should be retained by humans".[2] The UK, along with other actors, such as the United States, believe that a full ban on LAWS could be counterproductive, and that there are existing governance structures in place to provide appropriate oversight over the research, development, and deployment of automated weapons systems:

> …[T]he UK already operates a robust framework for ensuring that any new weapon or weapon system can be used legally under IHL. New weapons and weapons systems are conceived and created to fulfil a

> specific requirement and are tested for compliance with international law obligations at several stages of development.[3]

The UK is also interested in a "technology-agnostic" focus on human control because it believes that it will "enable particular attention to be paid to the key elements influencing legal, ethical and technical considerations of LAWS", as opposed to "debated definitions and characteristics" which, ultimately, may "never reach consensus". The position emphasises that taking a "human-centric, through-life" approach would enable human control to be considered at various stages and from multiple perspectives. This includes across all Defense Lines of Development, the acquisition of weapons systems, and their deployment and operation. It is the UK's position that the existing institutional infrastructure builds-in accountability measures throughout the weapon system lifecycle.

3. Methodology

In order to stress-test the governance and accountability structures that exist for UK weapon systems, and how they would apply to LAWS, we developed a hypothetical future scenario in which a UK LAWS kills a civilian during an extraction mission in Egypt. In order to ensure a level of feasibility and accuracy of construction, the scenario was built based on a wargaming scenario publicly published by RAND.[4] We then ran a facilitated role-play exercise based on our modified scenario with an initial group of Cambridge-based experts. With their feedback and the lessons from the role-play, we developed the final version of the scenario which we then used in the research study (see Appendix).

This final iteration of the LAWS scenario was used to run a mock UK parliamentary inquiry through which we interviewed 18 experts across various areas of expertise, including (but not limited to) UK military strategy, military procurement, weapons development, international humanitarian law, domestic military law, military ethics, and robotics.

The interviews ranged from 45 to 120 minutes and explored a variety of questions regarding the case. The main objective of the interviews was to catalyse a meaningful discussion around what information the experts deemed important and necessary in order to decide who should be held accountable in the aftermath of this scenario. A sample of the questions asked include:

- Who holds the burden of accountability and responsibility?
- What explanations and justifications for actions are needed?
- What information is necessary to come to a conclusion about the burden of accountability?
- Are there any foreseeable gaps because of the autonomy of the weapons systems?

The responses and dialogue of these 18 interviews were then reviewed and synthesized in order to develop a landscape of strengths and weaknesses of the current governance and accountability schemes related to UK institutions as they relate to LAWS, as well as recommendations on addressing any identified weaknesses. The full report is under preparation, but we are happy to share our preliminary key findings and recommendations below.

4. Key Findings

The main takeaway from the "inquiry", from both a legal and organisational standpoint, was that assessing accountability is in the details. This contrasts with what we perceive as a dominant narrative of "meaningful human control", which focuses mainly on human control, and the design of that interaction, at the point of final targeting action. The disconnect between the accountability across a weapon's lifetime and the focus on final targeting decision was observed throughout the various expert interviews. "Meaningful human control" has become the *idée fixe* of domestic and international conversations for regulation of LAWS but it disadvantageously provides a limited lens through which most experts and relevant personnel think about accountability.

To contrast this heavily focused narrative, the interviews have highlighted a whole range of intervention points, where humans are expected to, and should be supported in making decisions that enable legal, safe, and ethical weapon systems. These are arguably points that should be considered in "meaningful human control". These include, but are not limited to:

Establishment of military need:

- defining military necessity for research, development, and/or procurement; and
- choice of technological approach based on political and strategic motivations.

(*Main related stakeholders: UK MoD; UK Defense Equipment and Support (DE&S); private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics*)

Technical capabilities and design:

- trade-offs between general applicability and tailored, specific solutions with high efficacy and guarantees on performance;
- awareness, training, and foreseeability of contextual factors about intended use situations that may affect the performance of the weapon system; and
- documentation and communication of known limitations and failure modes of the system design.

(*Main related stakeholders: private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics; UK Defense Science and Technology, UK Defense and Security Analysis Division*)

Human-computer interaction design:

- choices of what data to include and what data to exclude;
- trade-offs between clarity and comprehensiveness;
- level of technical information communicated; and
- parallel communication channels: to operator in/on the loop, to command centres further from the field, and to logs for future technical analysis or legal investigation.

(*Main related stakeholders: Private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics; UK Defense Science and Technology; UK Defense and Security Analysis Division; UK MoD military personnel — human operators*)

Weapons testing:

- choice of parameters to be evaluated, frequency of evaluation, and conditions under which to evaluate;
- simulation of adversaries and unexpected situations in the evaluation phase;
- evaluation of HCI in extreme conditions; and
- evaluation of the human-machine team.

(*Main related stakeholders: private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics; UK DE&S; UK MoD military personnel — human operators*)

Procurement:

- robust Article 36 review;
- assessment of operational gaps, and trading-off operational capability with risks;
- trade-off between cost effectiveness and performance of weapons systems;
- documentation and communication of trade-offs so they can be re-evaluated as the context or technology changes;
- number and type of systems;
- provisioning of training and guidance; and
- provisioning for maintenance.

(*Main related stakeholders: UK DE&S; Article 36 convened expert assessment group; private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics*)

Weapons deployment:

- informing commanders about the capabilities and limitations of the system, of their track record in similar situations, and of novel parameters of the new situation;
- establishing and training for appropriate pre-deployment testing schemes to capture any vulnerabilities or "bugs" with specific weapons system;

- checking for readiness of troops to operate and maintain systems in the arena; and
- expected response of non-combatants to the presence of the weapon system.

(*Main related stakeholders: UK MoD commanding officers; UK MoD military personnel — human operators*)

Weapons engagement:

- awareness of limiting contextual factors, need to maintain operator awareness, and contextual knowledge; and
- handover of control between operators during an operation.

(*Main related stakeholders: UK MoD military personnel — human operators*)

Performance feedback:

- ensuring a meaningful feedback process to guarantee process improvement, reporting of faulty actions, communicating sub-par human-machine techniques and capabilities, and more.

(*Main related stakeholders: UK MoD military personnel — human operators; UK MoD commanding officers; UK DE&S; private military contracting companies, such as BAE Systems, Qinetiq, General Dynamics*)

5. Recommendations

5.1 Dialogue shift: Emphasising control chain and shared responsibility

The prioritisation of "meaningful human control" for LAWS-related risk mitigation and governance anchors the scope of control points around final targeting decisions. The narrative implies that this is the main area of control that we want to manage, focus, and improve on in order to ensure that the weapons systems we are deploying are still acting with the intent and direction of human operators. Although this is an important component of ensuring thoughtful and safe autonomous weapons systems, this is only a fraction of the scope of control points. In order for us to acknowledge the other points of control throughout the research, development, procurement, and deployment of LAWS, we need to be inclusive in our dialogue about these other points of human control.

5.2 Distribution of knowledge: Personnel training

Training everyone who touches the research, development, deployment, etc. of LAWS on international humanitarian law, robot ethics, legality of development, responsibility schemes, and more, would contribute to a more holistic approach to responsibility and accountability, and, at its best, can contribute to a culture that actively seeks to minimise and eliminate responsibility gaps through a collaborative governance system.[5] This distribution of understanding around governance could provide a better landscape for accountability through heightened understanding of how to contextualise technical decisions. Further, it can provide an effective, granular method for protecting against various levels of procedural deterioration. With shifting geopolitical pressures, as well as various financial incentives, there could easily be a deterioration of standards and best practices. A collaborative governance scheme that is based on a distributed understanding of standards, military scope, international norms, and more, can provide components of a meaningful and robust governance plan for LAWS. This distribution of knowledge, though, must be coupled with techniques for reporting and transparency of procedure to be effective.

5.3 Acknowledging the politics of technical decision-making/ design specifications

"Meaningful human control", through its dialogue anchoring, also puts a heavy burden on the technical components of design decisions, such as best practices for human-computer interactions. The politics of quantification in technical decision systems for autonomous systems should not be undervalued. The way any autonomous system decides what actions to take and what information to show is a highly political decision, especially in the context of war. It is important to understand which parts of the design process are more political than they are technical, who should be involved in those decisions, and how to account for those decisions in the scope of research and development (to inform a proper, comprehensive collective responsibility scheme).

Appendix available online at https://doi.org/10.11647/OBP.0360#resources

Notes and References

1 Carpenter, C. 'From "stop the robot wars!" to "ban killer robots"', *Lost Causes* (2014), pp. 88–121. https://doi.org/10.7591/cornell/9780801448850.003.0005

2 Human Machine Touchpoints. *The United Kingdom's Perspective on Human Control Over Weapon Development and Targeting Cycles* (2018).

3 Human Machine Touchpoints (2018).

4 Khalilzad, Z. and I. O. Lesser. *Selected Scenarios* from *Sources of Conflict in the 21st Century*. RAND (1998), pp. 317–18.

5 Ansell, C. *Collaborative Governance* (2012). https://oxfordindex.oup.com/view/10.1093/oxfordhb/9780199560530.013.0035

22. Representation of Future Generations in United Kingdom Policy-Making

Natalie Jones, Mark O'Brien and Thomas Ryan

Highlights:

- Global existential and Catastrophic Risks, particularly those arising from technological developments, present challenges for intergenerational justice. This chapter presents a solutions-based approach to this challenge by examining options for representing future generations in our present policy-making structures, drawing on case studies from Singapore, Finland, Hungary, Israel, Scotland and Wales.
- The authors derive several factors which contribute to the success of some of these institutions, and discuss reasons for the failure or abolition of others. They draw out broad lessons which can be applied to policy-making in the UK and use these to make a number of recommendations.
- At the policy level, legislation should be passed containing an obligation to include the long-term risks of any Government Bill in its Explanatory Note and intergenerational rights should be included in any potential British Bill of Rights, if and when this is passed.
- At the institutional level, an All-Party Parliamentary Group on Future Generations should be formed and the various futures

https://doi.org/10.11647/OBP.0360.22

research institutions and think tanks should cooperate to form an expert advisory panel with a mandate to influence policy.

- In the longer term, political momentum should be translated into a formal Select Committee on Future Generations.

This chapter considers a range of options for bringing long-term issues around the management of extreme global risks into existing democratic arrangements, as recommended by Chapter 2 and Chapter 17. Its lead author subsequently worked to establish an APPG for Future Generations (https://www.appgfuturegenerations.com), which in turn supported the development of a Welfare of Future Generations Bill (https://todayfortomorrow.org.uk). CSER continues to play an active role in supporting these initiatives and pushing for improved representation of future generations in policy-making.

1. Introduction

Global Catastrophic and existential risks pose central challenges for intergenerational justice and the structure of our current democracy. The Global Challenges Report 2016 defines Global Catastrophic Risk as risk of an "event or process that, were it to occur, would end the lives of approximately 10% or more of the global population, or do comparable damage".[1] A subset of catastrophic risks are "existential" risks, which would end human civilisation or lead to the extinction of humanity.[2] Catastrophic and existential risks may be categorised in terms of ongoing risks, which could potentially occur in any given year (e.g. nuclear war; pandemics), versus emerging risks which may be unlikely today but will become significantly more likely in the future (e.g. catastrophic climate change; risks stemming from emerging technologies). Ongoing risks have existed for some time now and are generally well understood. However, emerging risks, particularly those arising from technological developments, are less understood and demand increasing attention from scientists and policy-makers. These technological developments include advances in synthetic biology, geoengineering, distributed manufacturing and Artificial Intelligence (AI).[3] Although the impact of these technologies is still very uncertain, expert estimates suggest a non-negligible probability of catastrophic harm.

In this article we rely on two main premises. The first is that future generations are under-represented in current political structures partly due to political "short-termism" or "presentism".[4] Governments primarily focus on short-term concerns, which mean that they may systematically neglect Global Catastrophic Risks and, accordingly, future generations.[5] The problem of presentism transcends political divisions: people across the political spectrum are concerned about its effects, and should care about mitigating Global Catastrophic risks. This situation is exacerbated in that the good of mitigating Global Catastrophic and existential risks is typically global. Individual political actors (even whole countries) bear many costs in providing for such goods, whereas the benefits are dispersed globally. In addition to the benefits of mitigating existential risks being global, many of the beneficiaries are future people who do not exist presently and as such have no voice in the political process. There is a clear lack of incentives to mitigate such risks, and market failure should be expected.[6]

The second key assumption is that we as a society consider the rights and interests of future generations to be important. It is beyond the scope of this chapter to present a complete account of the philosophical arguments on this matter. It is sufficient to note that although significant philosophical problems have been pointed out, chiefly due to the fact that the actions of present people have a causal impact on the values, number and identity of future individuals,[7] there are several theories of intergenerational justice that may support this assumption.[8]

The need to include explicit pathways in governance structures for accountability to the rights and needs of future generations has been noted.[9] Some thought has been put into how future generations may be represented in relation to environmental risks such as climate change, resource depletion and biodiversity loss; this research is reflected in the sustainable development literature.[10] However, this problem has not been explored in relation to society's burgeoning awareness of technology-related catastrophic and existential risks. In addition, such pathways have not been fully explored in the United Kingdom (UK) context. This chapter hopes to fill this gap in the literature.

We aim to present a solutions-based approach to the challenge of intergenerational inequality. This chapter will examine options and challenges for representing future generations in our present policy-making structures. In practice, Wales and Scotland both have institutional forms of representation for future generations. We therefore focus here on England, while also considering options that could be mainstreamed throughout the UK.

In Part 2 of this chapter, we explore case studies of future generations representation from several different countries, including Singapore, Finland, Hungary, Israel, Scotland and Wales. We derive several factors which contribute to the success of these institutions, and discuss reasons for the decline of some. We draw out broad lessons which we can apply to policy-making here. We go on in Part 3 to discuss the specific UK policy context which may affect the appropriate solutions, and in Part 4 we explore policy options and make recommendations based on our previous findings. Present generations pose a much greater risk to future generations than any past generations posed to the present generation, due to a combination of fast economic growth and unprecedented scientific advancement and technological development.[11] The time is ripe for the futures studies and existential risk communities to connect with policy-makers on these important issues.

2. Institutional Case Studies of Representation of Future Generations

Over the last two decades, several national governments have set up institutional structures to attempt to address short-termism in decision-making, with varying levels of success. These institutions have taken a variety of different structural and functional forms, providing a useful data set by which we can analyse factors contributing to their success. Here we focus on institutions explicitly aimed at the interests of future generations, rather than those which may merely have an indirect effect on future generations (such as environmental protection agencies).

We discuss the main variables in institutions in terms of structure, function and degree of power. Structurally, commissioners and committees have been used, with varying amounts of resources at their disposal. The independence of such institutions from government has varied considerably, from taking the form of companies at arm's length from government, to being composed of Parliamentarians themselves. Similarly, the responsibilities and powers of each institution ranges from a minimalist research and advocacy role, to the power to delay or block legislation. Subject scope also varies, as do the individuals and organisations that institutions work with: in particular, we find that only one institution has explicitly considered Global Catastrophic or existential risks in its work. In addition, the historical and social context within which these institutions were created, and the accompanying political pressures, naturally differ among countries. We analyse these variations to determine if their successes can be transferred to the English context.

It is important for the purposes of our analysis to specify what indicators are being used to assess the success of these representative institutions. Broadly, one of the most important indicators of success for these institutions is the *impact* they have had on present decision-making to take intergenerational interests into account. Unfortunately, this indicator is necessarily vague; almost all institutions differ somewhat in their functions and powers, and giving a narrow definition of "impact" will wrongfully exclude institutions which take alternative measures to ensure present representation of future generations in decision-making. However, it will become clearer what kinds of impacts are desirable.

Another success indicator is *increasing dialogue and giving a clearer articulation of intergenerational issues* in the political and public spheres. Presently, as we have already seen, the issue of representing the rights and interests of future generations is not well articulated (if at all) in the UK political context. Simple awareness of these issues is an essential step towards their having an impact upon decision-making.

A third key indicator of success is *longevity*. A trend with intergenerational representation mechanisms is that such institutions generally face challenges to their status within a short period after their creation. But longevity is essential for successful representation of future interests.

The institutions discussed are summarised in Table 1, which shows dates of operations, position with respect to the executive and the legislature, scope, and powers.

Table 1: Institutions for representing future generations.

Country	Dates of operation	Position with respect to executive and legislature	Scope	Functions and Powers
Finland	1993-	Standing Committee of Parliament	Futures in general; can choose own scope	Research/advisory Education
Hungary	2008–2012	Structurally independent from government	Issues which may affect the constitutional right to a healthy environment	Research/advisory Complaints investigation Legal enforcement
Israel	2001–2006	Parliamentary committee	Environment, natural resources, science, development, education, health, state economy, demography, planning and building, quality of life, technology, law, any other matter considered relevant	Research/advisory Initiate legislation Veto legislation
Scotland	2005-	Structurally independent from government	Futures in general; can choose own scope	Research/advisory Education
Singapore	2009-	Within the Prime Minister's Office	Risk and futures; can choose own scope	Research/advisory Education
Wales	2016-	Structurally independent from government	Sustainable development	Research/advisory Recommendations are binding

2.1 Finland: Committee for the future

Created in 1993, the Committee for the Future is a Standing Permanent Committee of the Finnish Parliament. It consists of 17 Parliamentarians representing all parties, in proportion to the makeup of Parliament itself.[12] The Committee serves a variety of functions: it acts in a "think tank" role for government by analysing research regarding the future and assessing possible implications for the work of Parliament; it conducts dialogues with other organs of government on any foreseeable long terms issues affecting policy or the work of the bodies in question; it prepares responses to Government reports on the future of Finland which are commissioned by the Prime Minister every four years; and it engages in public outreach.[13] Aside from these reports, it is free to choose its own methodology and the scope of issues upon which to focus.[14] It is also responsible for and must cover the implications of technological development for society. Formally, the Committee has little power to intervene in legislation or policy decisions, and has no power to receive and act legally upon complaints from the general public.

Nevertheless, the Committee appears to have had substantial impact. It has demonstrated agenda-setting power in the Parliament, and the government has tended to adopt the Committee's responses to its reports.[15] The Committee is also the longest-running institution assessed in this analysis, which indicates that it has achieved a stable relationship and balance of power with government.

This success may be due to a number of factors. First, the Committee's work had legitimacy from the beginning due to widespread cross-party and public support during its creation.[16] At that time Finland already had a substantial history of futures studies, concentrated in the Finnish Society for Future Studies. The Committee's continuing public outreach work can only sustain this legitimacy. Secondly, the Committee has enough power to have an impact, whilst not enough power to provoke any major challenges to its status. Despite the lack of significant independence from government, it has been able to set its own agenda for the most part, meaning it can challenge a wide scope of issues which it sees as relevant to future generations. The fact it is composed of Parliamentarians allows the opportunity for informal intervention by its members, lends its findings political weight, and is a strength in that its proceedings are highly integrated with those of Parliament.[17]

2.2 Hungary: Commissioner for future generations

The Hungarian Commissioner for Future Generations was one of the strongest representative mechanisms for future generations yet created. The Commissioner was established in 2008, but only continued until 2012 before having its power substantially reduced.[18] Structurally, the Commissioner was elected by Parliament, but under the Act LIX of 1993 on the Parliamentary Commissioner for Civil Rights (Ombudsman) 1993 had to fulfil the condition of being a lawyer with expertise in environmental protection and/or nature conservation law (s 27/A. § (2)). Independence was also assured by the exclusion of anyone who had, among others, held office or been a member of a political party within the last 4 years, or held other employment or business that could constitute a conflict of interest (ss 3, 27/A(2)). Structurally, the Commissioner was elected by Parliament, but was required to be a lawyer with expertise in environmental protection and/or nature conservation law (s 27/A. § (2)). Independence was also assured by the exclusion of anyone who had, among other criteria, held office or been a member of a political party within the last four years, or held other employment or business that could constitute a conflict of interest (ss 3, 27/A(2)).

In terms of scope, the primary task of the Commissioner was to "ensure protection of the fundamental right to a healthy environment", which at the time was enshrined in Hungary's constitution. The Commissioner's core duty was to receive complaints and carry out investigations in relation to all issues that may affect citizens' constitutional right to a healthy environment (s 27/B). These investigations often resulted in legal cases taken by the Commissioner — over 200 substantive cases a year, many of which resulted in success.[19] Through this investigatory role it achieved many successes in protecting the interests of future generations.[20] In addition, the Commissioner was also responsible for strategic development research, and consulted on legislation concerning the environment and all levels of government. The Commissioner had considerable powers, including the power to call for termination of activity damaging the environment, backed up by police and law enforcement bodies.

Advantages of the Hungarian approach include that the office was legally (and arguably politically) independent from other government branches and from businesses, and had some legitimacy through its support from civil society groups and its interaction with individual citizens through its complaints service. It also maintained transparency and open relationships with all stakeholders during investigations and reported annually on its work (s 27/H). However, the Commissioner had fairly narrow scope, both in terms of its issue focus (i.e. environmental issues) and methodology; the Commissioner seems to have expended a great deal of resources on legal pursuits in response to individual complaints.

Additionally, the institution did not see the longevity essential for long-term representation of future interests. The role ended in 2011 when Hungary's four commissioners (on different subjects) were amalgamated into one position, the powers and mandate of the Commissioner were vastly reduced and it faced large budget cuts. This change was a part of a new constitution, drafted by the newly incumbent right-wing Fidesz party. It is likely that, given the Commissioner's notable interventions in private and governmental interests, there was significant political pressure to reduce its level of power. Despite the fact that originally, the Commissioner was brought about by support from across the political spectrum and from civil society groups, there may still have been a deficit of political understanding of, or sympathy for, its goals and methods. Whilst the Commissioner did engage with citizens through its complaints role, it may still have lacked the widespread awareness and support for tackling intergenerational issues to prevent it being easily dissolved by other political interests.

2.3 Singapore: Centre for strategic futures

The Centre for Strategic Futures (CSF) is an in-government futures think-tank established in 2009 within the Strategic Policy Office, which is itself a part of the Prime Minister's Office of Singapore. [21] Focusing on the public sector, CSF works to encourage and improve governmental and cross-department strategic thinking on risk and the future. This can be seen both in the wide audience it has reached through educational and networking methods within the civil service,[22] as well as through individual projects with other departments, such as that on the implications of automation on the Singapore workforce (carried out conjointly with the Minister of Manpower).[23]

Structurally, whilst its position within the Prime Minister's Office may lend it some authority in political and policy spheres, it also raises questions of independence. The precarious position it occupies close to government means it is open to both political pressures on agenda-setting and outright dissolution if it causes much upset for the relevant stakeholders. However, there is reason to think these latter worries do not pose much of a threat. Singapore has a history of valuing strategic thinking and scenario planning that dates back to the 1980s,[24] and as such, the relevance of the institution is firmly ingrained in the civil service and government. Furthermore, the head of civil service has written glowing reviews of the Centre's work in introductions to its annual report, "Foresight".[25]

Functionally, CSF acts mainly as a futures think-tank for government and the civil service. It has worked on a wide range of issues in doing this, including the effects of automation and renewable energies on Singapore, as well as more abstract questions of national identity.[26] Yet, its most distinctive feature lies in its role to, "not just to think about the future, but also to think about how we think about the future".[27] The Centre has developed highly rigorous frameworks for thinking about future trends, risks and opportunities. An example is its "Scenario Planning Plus" (SP+) toolkit, which incorporates insights from chaos theory on complex systems,[28] and psychological insights on cognitive biases when thinking about the future.[29] Furthermore, it has stressed the need to pick up on "weak signals" which might be evidence of upcoming, significant future events.[30] A major benefit of such a framework is its receptivity to low-probability, high-impact events, such as Global Catastrophic and existential risks.

CSF's second main role is to encourage and facilitate this thinking across policy-making platforms. In addition to encouraging individual departments to engage in strategic thinking about the future, the Centre aims to facilitate wider, "whole-of-government" thinking and coordination on future issues, which is advantageous since long-term risks and opportunities do not all necessarily fall into neat public service categories.[31] It has partly achieved this through running "Futurecraft" workshops to teach its SP+ toolkit to members of the civil service, and trainees of the Civil Service College.[32] This outreach, along with the annual publication of its Foresight reports, means the Centre increases transparency and is accessible to individuals across the public sector.

CSF lacks any substantive powers to intervene in the legislative process, or penalise those which it sees as acting against the long-term interests of Singapore. However, this has not been an issue given its role in promoting long-term, strategic thinking, which mainly requires positive action on its part. Furthermore, although the Centre has not engaged in extensive outreach work with the general population of Singapore, it has made efforts to engage with relevant professionals from a range of backgrounds "through incoming visits, overseas trips, paid consultancies, interviews and curated events".[33]

As an institute for implicitly representing future interests, CSF has been broadly successful and has several key, desirable features such as its focus on inculcating strategic thinking on the future across government to disperse its workload and enhance scope.

However, several features of the Singaporean context mean that this institution may not be easily transferable to the UK. First, Singaporean politics arguably does not suffer from political short-termism to the same extent. Partly as a consequence of the design of the Parliamentary system, the ruling People's Action Party has been in power for half a century. Although individual Parliamentarians are at risk of losing their seats, there is not enough of a threat to undermine the government planning far into the long-term. The government has acted favourably towards strategic future thinking since the 1980s, and there are little signs that it will change path in the near future.

Secondly, there are factors intrinsic to Singapore as a nation which dispose it to allocate more resources to long-term planning. Its relative youth as a nation (having only achieved full independence in 1965) as well as its precarious location, size and lack of natural resources gives rise to feelings of national insecurity (similar factors likely influenced the creation of Israel's Commission for Future Generations). Furthermore, arguably an increased cultural emphasis on collectivism and national prosperity, and diminished value placed on individual freedom, creates a context more favourable to long-term planning and strategy.

2.4 Israel: Commission for future generations[34]

Established in 2001 by the Knesset (Israel's Parliament), the Israel Commission for Future Generations was an organ of Parliament headed by a Commissioner chosen by an ad-hoc Parliamentary committee and appointed by the Speaker of the Parliament.[35] Similarly to Hungary, regarding independence, the Commissioner could not be someone whom in the last two years had been active in political life or a member of any political party. The Commissioner was assisted in its role by a Public Council (an advisory committee) which consisted of scientists, intellectuals, clergymen and other public figures. The Commission is now disestablished; it was only given a five-year mandate and when the term of the first Commissioner ended no new Commissioner was appointed, apparently for budgetary reasons.[36]

Functionally, the Commissioner could give opinions on bills and secondary legislation brought before Parliament if they believed it concerned future generations. It also had the power to initiate bills to advance the interests of future generations, and could play a general advocacy role to Parliament and Parliamentarians. It was required to submit an annual report on its activities for that year, creating some transparency.

The scope of its responsibilities was wide, stretching across 12 policy areas including environment, development, science, and technology. Furthermore, the explanatory notes to the Knesset Law explicitly contemplated the possibility of adverse consequences from genetic engineering or other technological developments. This is the closest reference to existential risks across any of the institutions being assessed.

As well as holding the power to initiate bills in the Knesset, the Commissioner had an effective veto power over the passage of legislation which didn't comply with the interests of future generations. This may be one of the reasons the institution was eventually scrapped: alongside cost issues, members of the Knesset cited "their feelings that the Commission received too much authority to interfere in their work".[37]

2.5 Scotland: Future Forum

Set up by the Scottish Parliament in 2005 as a company at arm's length from the Parliament itself, the main motivation for the Future Forum was to tackle short-termism in present decision making: to "look beyond immediate horizons, to some of the challenges and opportunities we will face in the future".[38] A Board of Directors helps guide the Forum's work; its members include backbench MSPs (Scottish Parliamentarians), prominent academic leaders, civil servants and business leaders. The Forum is autonomous from the Parliament in deciding the focus of its work, though it still depends on it for funding.[39]

One of the main functions of the Forum has been to "stimulate public debate in Scotland" with respect to preparing for the future.[40] In doing so, it has engaged with politicians, the private sector, and the public. It also carries out "futures studies", reporting on how various areas of Scotland will evolve in the future.

In terms of success, the institution is laudable for making an active effort to directly promote longer term thinking in decision making. From 2011–2016, the Forum organised more than 100 events directed to bringing "'fresh-thinking' into the [Scottish] Parliament".[41] However, it is hard to assess the impact of these educational events on policy making in general. Furthermore, the Forum has thus far been limited in scope, dealing with only a handful of varied individual topics in its future studies research. This narrow scope is possibly affected by limited powers that Scottish Parliament has to deal with issues relating to economic policy, healthcare budget or existential risk research, and highlights the need for the UK Parliament to deal with intergenerational issues.

2.6 Wales: Commissioner for Future Generations

The Commissioner for Future Generations is a guardian role focused on sustainable development, outlined in the Well-Being of Future Generations (Wales) Act 2015. This is the most recent institution considered here: the first Commissioner came into existence on February 1, 2016. The Act imposes certain obligations regarding sustainable development and well-being targets on 44 listed Welsh public bodies, and the Commissioner's main role is to ensure that this is done successfully.

The Commissioner may research how public bodies can best meet these targets, as well as encourage and give recommendations to these bodies. The Act obliges public bodies to follow these recommendations, and the Commissioner can carry out reviews at their own discretion to assess their progress. In a wider role, the current Commissioner has emphasised the need for public bodies to engage with citizens on discussions of the future of Wales.[42]

It is too early to assess the success of the Commissioner given the institution's youth. Whilst it is promising to see long-term thinking being promoted across public bodies, it does not seem that Global Catastrophic and existential risks are being considered. Again, some issues may also not receive attention to their long-term consequences due to a lack of devolved power on Wales' part.

Why does the Commissioner exist in Wales but not England? What distinguishes the Welsh case? First, in Wales there is a more prominent strand of environmental and social awareness than in mainstream UK politics, and an element of "conscious exceptionalism" which made Welsh politicians enthusiastic to distinguish themselves from English MPs by adopting a sustainability agenda.[43]

In addition, Welsh environmental policy contains a strong emphasis on "management and stewardship" in environmental policy — that is, a policy context which foregrounds waste reduction and renewable energy.[44] In England, by contrast there is a much greater focus on three prominent short-term issues: flooding, overcrowding,[45] and coastal erosion. These issues are important, but do not provide as strong a platform for intergenerational sustainability because they inherently respond to short-term complaints such as housing. It is useful to observe that the future generations agenda had cross-party support in the Welsh assembly, and secondly that the UK government's disbanding of the Sustainable Development Commission — expanded upon in the next section — "created a shared understanding of the fragility of a purely administrative structure, not backed by legislation".[46]

2.7 Conclusions

Representative institutions for future generations, whether local or abroad, differ widely in their structure, functions and power. Although such institutions have only begun to appear in the last two decades, common trends and features exist. In particular, they tend to face challenges to their existence within a few years of their creation (usually an election cycle). This is a major problem for securing successful representation of future generations. The representation mechanisms that we propose will therefore seek to avoid capricious party politics, either by being firmly constitutionally entrenched, or more realistically by being a cross-partisan organ that recognises its limits and works with the political grain. As such, several factors can be drawn out from the analysis which may increase or decrease the likelihood of short-term discontinuation of a future representative institution.

First, institutions which are given too much power, too early in their lifespan, tend to face rejection from politicians. The Israeli and Hungarian Commissioners illustrate this pitfall. This is a difficult balancing act, however: an institution with no power is of no use in representing future generations. But the sort of massive, transformational change needed to protect future generations requires a degree of institutional strength — strength which appears to be deeply incompatible with current politics. This implies a major dilemma — a choice between proposals which are ineffective in protecting future generations but politically realistic, and those which are effective yet unrealistic — which will be returned to in our conclusions.

The legitimacy of, and public support behind, an institution is a key factor as to whether it will last. Public and politicians alike need to perceive an institution as legitimate, and its functions and powers must be proportional to this perceived legitimacy. Public and political (especially cross-party political) support for future representative institutions is essential for representative institutions to have any level of power. It is imperative, then, for any such institution to be transparent and accessible in its work, as well as taking initiative to promote the cause of intergenerational rights and issues to the general public and decision-makers. Civil society movements and support can be very advantageous in the success of implementing long-term thinking in policy (the creation of the Hungarian Commissioner due to this is illustrative). Public and political engagement of the cause is key to successfully representing future generations in the long-term.

Structurally, it has been beneficial to have a multi-disciplinary team working on the issues, as in Scotland and Hungary. This makes sense given the wide range of issues affecting future generations. Furthermore, securing the right kind of independence from government is key to ensure criticisms of policy can be made without fear of dissolution, as well as to maximise impact. Although inclusion of Parliamentarians can risk a conflict of interests, their participation lends political weight to the institution, both in terms of influence and the importance of the institution. This may be essential for the highly influential, long-lived Finland Committee. It is also important to ensure independence in agenda setting, at least to an extent, as observed in Finland and Scotland. However, academic engagement should be used to prioritise issues. In making these findings we echo the argument of the World Future Council that the key characteristics of future-representation institutions should be independence, transparency, legitimacy, access to information, accessibility, and authority.[47]

We are led to the preliminary conclusion that in the UK Parliamentary context, substantive powers should not be given to intergenerational representatives, at least initially (contrast the Israeli power to veto legislation, and the Hungarian abilities to enforce rulings). Instead, a UK-wide representative institution could play a monitoring role for legislation affecting future generations; carry out and collate relevant research with respect to intergenerational issues; play an advisory role to government; and work to create wider public awareness of intergenerational inequality issues.

3. English Policy Context

Several England-specific factors are important in determining which policy options should be adopted in order to mainstream the representation of future generations.

3.1 UK Sustainable Development Commission

The UK previously had a Sustainable Development Commission (SDC), responsible for promoting sustainable development throughout the UK.[48] The SDC reported to the UK government, providing analysis of government departments' Sustainable Development Action Plans and responding to consultation papers which often disagreed with Government Policy.[49] The Commission's work on sustainable development was relevant to future generations, though they were not its explicit mandate.

However, the Commission was not statutorily independent, which may have limited it in its criticism of government policy, and also enabled the government to easily remove it in 2010.[50] Although the exact motivations for this are unclear, it seems likely that targeted criticism of government actions may have had an impact.

3.2 Environmental Audit Committee

Historically, the Environmental Audit Select Committee monitored the sustainability policies of government departments in a way similar to that prescribed by the Welsh Act. This is an important precedent in any attempts to introduce future representation into government, in particular because the Committee in 2011 recommended the creation of a new cabinet minister for sustainable development,[51] in order to improve the situation post-abolition of the SDC. Our concerns are broader than this: future representation encompasses a large range of discrete concerns than sustainable development. However, this is a useful recommendation which may be updated according to our understanding of intergenerational justice.

3.3 Political discourse regarding future generations

Political discourse in the UK places a strong emphasis on responsibility to future generations. This is reflected, for instance, in the political discourse surrounding national debt and austerity since 2010, which revolves around ideas of what today's voters owe to future UK citizens. The idea that each generation should "live within its means" has gained support even from the radical political opposition[52] and has been explicitly linked to intergenerational equity by the Prime Minister.[53] Potential exists to ground policies regarding representation of future generations in already existing concepts in British public discourse.

3.4 Merger of the Department of Energy and Climate Change

In 2016 a restructuring of government departments led to the merger of the former Department for Business, Innovation and Skills and the Department of Energy and Climate Change. The latter once dealt with many of the sustainability issues that have historically been at the heart of the intergenerational justice movement. This change has been understood by some as a signal that the government is not committed to sustainable action on climate change, although the government disagrees.[54] The new Department for Business, Energy and Industrial Strategy retains a minister for climate change, and some have argued that it may constitute a better foundation for the decarbonisation of the British economy.[55] If criticisms of the merger are correct, then this may indicate that the political environment is not supportive of future-planning and issues of intergenerational inequality.

3.5 The UK Constitution

The structure of constitutional law creates a distinct challenge to any attempt to institutionalise representation of future generations in England. In states like Hungary, future commissions can be created by constitutional law and protected against governments who must then rely on sweeping change if they wish to remove them. In the UK, on the other hand, no laws are more fundamental than any others; any statute can simply be repealed by Parliament. In addition, a key constitutional principle is that Parliament may not bind itself for perpetuity. A number of pieces of legislation have attempted to introduce a longer-term view, with various levels of success, such as the Human Rights Act 1998, the Climate Change Act 2008, and the Fiscal Responsibility Act 2010. These demonstrate the possibility of overcoming constitutional challenges.

Due to the uncodified and organic character of the British constitution, Parliamentary politics are governed by convention to a relatively large degree. These conventions are more helpful to the implementation of future representation in the UK, because they (as opposed to explicit documents) will dictate the tools and avenues of institutional form, and in some cases because they may create or enable an institutional resistance to change — particularly a change as large as a general perspective shift toward the future.

4. Recommendations

On the basis of the comparative analysis and UK policy context presented above, we make several recommendations. Each recommendation is followed by a brief explanation.

To begin, we note the following caveat. As noted earlier, there is a certain dilemma in that, globally, futures institutions with more power than politically acceptable have been quickly abolished, while those which are politically tolerable are not powerful enough to make the kinds of truly transformational changes required to protect future generations. In the face of this dilemma, we have chosen proposals which are practicable and politically feasible, taking the view that a small step forward is better than no step at all. We acknowledge the criticism that these proposals may not be nearly enough, but note that they may provide a foundation from which more radical change can be sought.

a) An All-Party Parliamentary Group on Future Generations should be formed.[56]

All-party Parliamentary Groups (APPGs) are multi-party groups of MPs who meet regularly or semi-regularly to discuss issues of common interest. They are registered formally in Parliament and are required to hold annual elections, but otherwise are informal groups organised by the interests of MPs for the sake of promoting particular causes. APPGs draw together members of major parties in order to maximise the possibility of influencing government. They create and enhance cross-party support, and as such we think that they are a good first step towards creating cross-party support for future generations issues. During their meetings, they discuss the activities of the governing parties and issues relevant to their subject of concern, and enlist government ministers to speak on their issues of concern. An APPG can use the existence of party members who deviate from the partisan line in order to give the issue in question greater exposure and to introduce it into legitimate party discourse.

An APPG may be a useful stepping stone to eventual institutionalisation of intergenerational justice in Parliament (perhaps in the form of a Select Committee). APPGs serve to increase the visibility of particular issues and emphasise their bipartisan support, creating a sense of the issue or "constructing" it as a shared, objective one. Another important function of APPGs is to act as a channel through which charities, campaign groups, NGOs and even commercial interests can involve themselves in government and political lobbying. This means that an APPG for future generations could function as a means by which the prominent civil society movement for sustainable futures could be translated into political change.

In practical terms, the lack of an explicit precedent for the representation of future generations in the British Parliament does not eliminate the possibility of an APPG for future generations. Many APPGs begin with the support of a prominent charity or other NGO, and perhaps the Centre for the Study of Existential Risk or the Future of Humanity Institute could operate as such a support in this case.[57] This would be an alliance reminiscent of that between Finland's Committee for the Future and the Finland Futures Research Centre in Turku University. Though such support is not a requirement, in practice an APPG needs some form of support in order to do its work effectively.

b) Legislation should be passed containing an obligation to include the long-term risks of any Government Bill in the accompanying Explanatory Note.

We recommend an obligation to describe the long-term risks of any bill introduced into Parliament, and to include this in the accompanying Explanatory Notes. Micro-level measures such as this are somewhat outside the scope of this chapter, which focuses on institutionalised representation, and we include this as just one example. Further research should be done into other possible options to promote good risk management on the micro-level.

c) The various futures research institutions and think-tanks should cooperate to form an expert advisory panel with a mandate to influence policy.

There are several academic institutes and think-tanks in the UK which study catastrophic and existential risks, sustainable development, and the future of society. These include, but are not limited to, the Centre for the Study of Existential Risk, the Future of Humanity Institute, the Oxford Martin School, Forum for the Future, the Centre for Future Studies and the Intergenerational Foundation. A veritable wealth of expertise is contained here, and these institutions should consider working together to create a committee tasked with providing advice to government. This is a recommendation which would not require much immediate action from government, save a willingness to receive advice.

This sort of independent expert advisory group could be formalised in the form of a non-departmental public body (NDPB), which operates at arm's length from government. There is clear precedent here, for instance in the form of the Committee on Radioactive Waste Management, which is an NDPB. Another option would be for a Policy Advisory Group (PAG) to be formed, which is simply a panel of people who advise on policy development.

d) If and when a British Bill of Rights is passed, the opportunity should be taken to include intergenerational rights.

Institutional security is difficult to acquire in the UK government, exemplified by the case of the Sustainable Development Commission. One of the best opportunities to constitutionally secure rights for future generations may be in the currently proposed "British Bill of Rights". Such rights-focused statutes are typically politically difficult to repeal. If intergenerational justice becomes part of the lexis of codified "British rights", it may have acquired a foothold of such historical significance that repeal would later become a practical impossibility. However, more research would be needed on the precise legal formulation and content of such rights.

e) In the longer term, political momentum should be translated into a formal Joint Committee on Future Generations.

A joint committee should be formed, charged with scrutinising every government bill for its compatibility with the rights and interests of future generations, and investigating the extent to which government departments consider future generations in their operation. A joint committee, unlike a select committee, is made up of both MPs and Members of the House of Lords. This committee would be modelled on the Joint Committee on Human Rights, which is charged with scrutinising every government bill for its compatibility with human rights, and the UK's compliance with its international human rights obligations. This is a less immediate option than an APPG (which requires a mere 10 interested Parliamentarians to come together), as a joint committee needs to be created by Parliament via its standing orders and therefore necessitates a more involved process. However, in the medium term a joint committee would have more power than an APPG and would be a more effective way of representing future generations.

As an alternative option, it is important to note that select committees can appoint sub-committees to produce reports on particular issues.[58] In the future generations context, the select committee on Energy and Climate Change could be an appropriate candidate to appoint such a sub-committee.

f) Any Future Generations institution should be explicitly mandated to consider existential risks arising from technological development, in addition to environmental sustainability.

As previously noted, only the Israeli institution amongst our examples was mandated to consider risks arising from technological development; the other institutions only considered environmental risks. In light of the burgeoning research in this field demonstrating that technological risks are a serious issue, any institution mandated to address international inequality should expressly consider them.

g) Civil society needs to mobilise to form a strong cross-party support for representation of future generations.

A common factor amongst the successful institutions studied is that all were established against a background of significant support from civil society. In addition, in the cases where that support continued, and where civil society organisations created significant public awareness of future generations issues, the institutions were more likely to endure rather than being abolished as soon as they fell out of political favour. Civil society needs to mobilise to form a strong cross-party support for the policy measures listed here.

4.1 Proposals we considered but do not recommend

We encountered several ideas which we do not include above, for various reasons. One of these is the proposal for a "third house of Parliament", or "Guardians", made by Rupert Read.[59] Under Read's proposal these Guardians, appointed randomly amongst citizens on the same principle as juries, would have the power to (a) veto new legislation that threatened the basic needs and fundamental interests of future people, and (b) force a review of any existing legislation that threatens such needs and interests. He also suggests similar structures within local governments. As we found previously, institutions with veto powers did not last long, and as such we do not think this "third house of Parliament" would be workable. In addition, we share concerns raised by Michael Bartlet about the proposed method of selection by lot.[60]

A second idea was an annual, designated day on which the House of Commons would discuss future generations issues. By analogy, events are held annually in the House of Commons for Human Rights Day and International Women's Day. We did not recommend this because although this might serve to publicise future generations issues, this kind of tokenisation of the rights and interests of future generations could create complacency and ultimately undermine the long-term, year-round work which needs to be done.

Another alternative way to represent future generations could be through a Royal Commission. A Royal Commission is an *ad hoc* advisory committee appointed by the government, in the name of the Crown, for a specific investigatory and/or advisory purpose. They generally exist for a limited time, on average taking between two and four years to produce a report, and have had a mixed impact. The work of the Royal Commission on Environmental Pollution spanned 40 years and had considerable influence,[61] but other commissions have had less impact or have even been disestablished before reporting.[62] We do not think a Royal Commission would be an appropriate means of representing future generations primarily because a Commission is generally time-limited and addresses a specific issue. The interests of future generations do not support such a “one-time” approach.

5. Conclusions

In response to the issues of intergenerational inequality raised by catastrophic and existential risks, we have presented several concrete options to represent future generations in current policy-making, founded on a comparative analysis of similar representative mechanisms worldwide. There are several limitations to what we have presented here. First, because our scope is necessarily limited, we deal only with “macro” mechanisms; we do not consider more specific legislative proposals in detail. Second, these conclusions are quite specific to the United Kingdom, and particularly the English context. In particular, the cultural context surrounding intergenerational issues may significantly differ between societies. Further research is needed to determine appropriate representative mechanisms in other countries, for catastrophic risks are a global problem and intergenerational inequality cannot be addressed only by one country acting alone. Finally, there is the dilemma previously mentioned: are all of these recommendations insufficient to truly protect future generations? Is it politically impossible to avoid irrevocable damage to future generations? Perhaps. We consider that *some* representation is better than *none*. Further, we do not wish to rule anything out, nor to lose hope. Future generations need us to keep on. We hope that the examples set by the six countries analysed here will be taken up across the globe.

Notes and References

1 Global Challenges Foundation and Global Priorities Project. *Global Catastrophic Risks* (2016).

2 Global Challenges Foundation and Global Priorities Project (2016).

3 Global Priorities Project. *Policy Brief: Unprecedented Technological Risks*. Future of Humanity Institute, Oxford Martin School, and Centre for the Study of Existential Risk (2014).

4 Thompson, D. F. 'Representing future generations: political presentism and democratic trusteeship', *Critical Review of International and Political Philosophy, 13*(1) (2010): 17. https://doi.org/10.1080/13698230903326232

5 Global Priorities Project (2014).

6 Beckstead, N. 'Unprecedented technological risks, *Future of Humanity Institute* (2013). https://www.fhi.ox.ac.uk/wp-content/uploads/Unprecedented-Technological-Risks.pdf

7 Parfit, D. *Reasons and Persons* (1st edition). Clarendon Press (1984).

8 Gosseries, A. 'Theories of intergenerational justice: A synopsis', *S.A.P.I.E.N.S, 1*(1) (2008): 61–71. https://doi.org/10.5194/sapiens-1-39-2008

9 Global Priorities Project (2014).

10 Brown Weiss, E. 'In fairness to future generations', *Environment, 32*(3) (1990): 6.

11 Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Publishing (2014).

12 Parliament of Finland. *Enacting Legislation: Committees*. The Parliament of Finland (n.d.). https://www.eduskunta.fi/EN/lakiensaataminen/valiokunnat/Pages/default.aspx

13 Groomsbridge, B. 'Parliament and the future: Learning from Finland', *The Political Quarterly, 7*(2) (2006): 273.

14 Association of Secretaries General of Parliaments. *CPI Review No. 188 Geneva* (2004), pp. 5–17, p. 12, para 4. http://www.asgp.co/sites/default/files/CPI_188.pdf

15 Association of Secretaries General of Parliaments (2004).

16 Groomsbridge (2006).

17 Groomsbridge (2006).

18 Future Policy. *The Hungarian Parliamentary Commissioner for Future Generations*. Future Policy (n.d.). http://www.futurepolicy.org/guardians/hungarian-parliamentary-commissioner/

19 Future Policy (n.d.).

20 Institute for European Environmental Policy. *Establishing an EU 'Guardian for Future Generations'* (2015), p. 5. http://www.ieep.eu/assets/1849/IEEP_2015_Establishing_an_EU_Guardian_for_Future_Generations_.pdf

21 Centre for Strategic Futures. *History*. CSF (2015a). http://www.csf.gov.sg/about-us/history

22 Centre for Strategic Futures. *Futurecraft*. CSF (2015b). http://www.csf.gov.sg/our-

work/futurecraft

23 Centre for Strategic Futures. *Foresight 2015*. CSF (2015c). http://www.csf.gov.sg/docs/default-source/default-document-library/csf-report-2015.pdf

24 Centre for Strategic Futures (2015a).

25 Centre for Strategic Futures (2015c).

26 Centre for Strategic Futures. *Publications*. CSF (2016). http://www.csf.gov.sg/our-work/publications

27 Centre for Strategic Futures. *Foresight 2015*. CSF (2015). http://www.csf.gov.sg/docs/default-source/default-document-library/csf-report-2014.pdf

28 Centre for Strategic Futures. *Our Approach*. CSF (2015d). http://www.csf.gov.sg/our-work/our-approach

29 Ho, P. 'Thinking about the future: What the public service can do', *Ethos, Cognitive Biases* (vol. 7) (2010). https://www.cscollege.gov.sg/Knowledge/Ethos/Issue%207%20Jan%202010/Pages/Thinking-About-the-Future-What-the-Public-Service-Can-Do.aspx

30 Centre for Strategic Futures (2015c).

31 Ho (2010).

32 Centre for Strategic Futures (2015b).

33 Centre for Strategic Futures. *Programmes*. CSF (2015e). http://www.csf.gov.sg/networks/programmes

34 The Knesset. 'Commission for future generations', *The Knesset* (n.d.). https://www.worldfuturecouncil.org/file/2016/10/Knesset-Paper.pdf

35 The Knesset (n.d).

36 Göpel, M. and C. Pearce. 'Guarding our future: How to include future generations in policy making', *World Future Council* (2013), p. 7. http://www.futurejustice.org/wp-content/uploads/2013/04/Ombudspersons_for_Future_Generations_Broshure_WFC.pdf

37 Teschner, N. *Official Bodies That Deal With the Needs of Future Generations and Sustainable Development*. Knesset Research and Information Center (2013), p. 3. https://www.knesset.gov.il/mmm/data/pdf/me03194.pdf

38 Wilson, S. *Scotland's Future Forum: A Guide and Legacy Report for the 2011-2016 Parliamentary Session*. Scotland's Future Forum (2016). https://spark.adobe.com/page/6fofj/

39 Wilson (2016).

40 Wilson (2016).

41 Wilson (2016).

42 Office of the Future Generations Commissioner for Wales. *Future Generations Commissioner Challenges Public Services to Build a Genuine Conversation With the Public*. Office of the Future Generations Commissioner for Wales (2016). http://futuregenerations.wales/wp/2016/11/03/future-generations-commissioner-challenges-public-services-build-genuine-conversation-public/

43 Institute for European Environmental Policy (2015).

44 Ogwr, O. B. 'Differences between England and Wales – The Higgit Question', *Oggy Bloggy Ogwr* (web log post) (2011). http://www.oggybloggyogwr.com/2011/08/

differences-between-wales-and-england.html

45 Dangerfield, A. 'London's housing crisis: Five controversial solutions', *BBC News* (2014). http://www.bbc.co.uk/news/uk-england-london-28377740

46 Institute for European Environmental Policy (2015).

47 World Future Council. *Guarding Our Future: How to Include Future Generations in Policy Making* (2014). http://www.futurejustice.org/wp-content/uploads/2014/06/brochure_guarding_en_final_links2.pdf

48 Sustainable Development Commission. *Framework Document on the SDC*. Sustainable Development Commission (2009), p. 4. http://www.sd-commission.org.uk/data/files/publications/20091113%20SDCFrameworkDocument.pdf

49 E.g. Jackson, T. *Prosperity Without Growth*? Sustainable Development Commission (2009). http://www.sd-commission.org.uk/data/files/publications/prosperity_without_growth_report.pdf

50 Vaughan, A. 'Government axes UK sustainability watchdog', *The Guardian* (22nd July 2010). https://www.theguardian.com/environment/2010/jul/22/government-axes-sustainability-watchdog

51 Saltmarsh, N. 'Environmental Audit Committee recommends minister for sustainable development', *Sustainable Development in Government*. National Archives database (10th January 2011). http://webarchive.nationalarchives.gov.uk/20120913014812/http://sd.defra.gov.uk/2011/01/environmental-audit-committee-recommends-minister-for-sustainable-development/

52 Watt, N. and P. Wintour. 'John McDonnell: Labour will match Osborne and live within our means', *The Guardian* (2017). https://www.theguardian.com/politics/2015/sep/25/johnmcdonnell-labour-will-match-osborne-and-live-within-our-means

53 Thomas, N. 'Theresa May defends Tories' economic record', Ft.com (2016). https://www.ft.com/content/f6610b6e-ea50-317b-bd40-ef587d1c6618

54 Rincon, P. 'Government axes climate department', *BBC News* (2017) http://www.bbc.co.uk/news/science-environment-36788162

55 Fankhauser, S. *Why the End of DECC Could Be Good News on Climate Change*. Grantham Research Institute on Climate Change and the Environment (2016). http://www.lse.ac.uk/GranthamInstitute/news/why-the-end-of-decc-could-be-good-news-on-climate-change/

56 This first recommendation was taken up: following the writing of this chapter, the All-Party Parliamentary Group for Future Generations was registered in October 2017, and the authors were involved in the process of its creation. Nevertheless, it is worth explaining here why such a move matters.

57 In the recently established APPG, the Centre for the Study of Existential Risk hosts the APPG's Secretariat.

58 Maer, L. and M. Sandford. *Select Committees Under Scrutiny* (1st edition). UCL Constitution Unit (2004). https://www.ucl.ac.uk/political-science/publications/unit-publications/111.pdf<i

59 Read, R. *Guardians of the Future: A Constitutional Case for Representing and Protecting Future People*, Green House (2012). http://www.greenhousethinktank.org/uploads/4/8/3/2/48324387/guardians_inside_final.pdf

60 Bartlet, M. 'How can Britain protect the interests of future generations?', *Open*

Democracy (2012). https://www.opendemocracy.net/ourkingdom/michael-bartlet/how-can-britain-protect-interests-of-future-generations

61 Owens, S. 'Experts and the environment — The UK Royal Commission on Environmental Pollution 1970–2011', *Journal of Environmental Law, 24*(1) (2012): 1–22. https://doi.org/10.1093/jel/eqr031

62 Institute for Government. *The Lost World of Royal Commissions* (n.d.). https://www.instituteforgovernment.org.uk/blog/lost-world-royal-commissions

23. Financing Our Final Hour

Luke Kemp, Haydn Belfield, Ellen Quigley, Julius Weitzdörfer and SJ Beard

Highlights:

- This chapter asks the question: “Should institutional investors act to reduce Global Catastrophic Risks? What are their obligations? And if they should, how can they best do so?” The authors review existing investor campaigns on global risks alongside literature from finance, economics, corporate law and ethics to identify and assess the different motivations and tactics for institutional investors to act on potential global catastrophic risks.
- There are at least four rationales that campaigns have already articulated: ethical arguments, legal considerations, financial incentives, and risk avoidance. To these the authors add two new justifications: long-term self-interest and universal ownership.
- Global Catastrophic Risks are relevant to corporate governance, and there are grounds for potentially generalisable ethical and legal obligations towards catastrophic risks. To this end, the chapter argues that a rational and ethical institutional investor should adopt a *Financial Hippocratic Oath* to not contribute to global risks, while those with deeper stakes, such as Universal Owners, should commit to a *Financial Oath of Maimonides*: to use their investments to minimise global risks.

https://doi.org/10.11647/OBP.0360.23

- The authors also ask how institutional investors can use their money and influence to reduce global risks. They propose six different tactics for achieving this end: contest, protest, request, divest, reinvest and acquest.
- Investor campaigns benefit from an issue with a clear, compelling "moral villain" (an actor that plausibly increases global risks), profitable alternatives for reinvestment, investor leverage and a strong underpinning campaign. These factors apply to many global risks, including climate change, nuclear weapons, and Lethal Autonomous Weapons (LAWs).

This chapter was written over many years as a response to issues raised by the global divestment movement. This work led, in part, to the University of Cambridge deciding to divest its own investments from fossil fuels via the role played by Dr Quigley on a secondment to the university's Chief Financial Officer. The role of political economy in extreme global risk is discussed in Chapter 2, while the consideration of different mechanisms for global risk reduction draws on the research presented in Chapter 19.

Introduction

In 2015 students at Swarthmore College crowded the halls of the administration building. They sang, picketed, and occupied it for 32 consecutive days. The sit-in protest was part of a campaign that dated back to 2011.[1] It originally demanded that the faculty withdraw their investments from the "sordid sixteen": a group of 16 US oil, gas and coal companies with appalling human rights and environmental track records.

This was the beginning of the "fossil free" divestment movement (divestment usually referring to investors' sale of shares (public equity) in target companies; it may also extend to the sale of bonds and other financial instruments). It has spread like a contagion since then. As of 2022, 1508 institutions, collectively worth US $40.43 trillion, had pledged to divest from fossil fuels.[2] The types of commitments vary. Some restrict their divestments to a particular type of fossil fuel, such as coal, while many direct their divestment commitments towards the top 200 fossil fuel companies by reserves held. The approach of the "fossil free" divestment movement has since been adopted by the International Campaign to Abolish Nuclear

Weapons (ICAN) through their '*Don't Bank on the Bomb*' campaign,[3] a group which pushes for institutional investors to divest from nuclear weapons producers and publishes annual reports (since 2013) on who produces and funds nuclear weapons (institutional investors are large entities that pool money to purchase different investment assets; these include banks, pensions, endowment funds, hedge funds and mutual funds).

Many other global risks exist or are on the horizon but have yet to be the subject of investor campaigns. "Global risks" are risks that profoundly threaten the global economy and society.[4] Such risks are increasingly well known, but surveys suggest that there is a gap between the higher concerns of scientists and those of business leaders.[5] Of these risks we are concerned with those that could produce a global catastrophe, ranging from killing a significant (10–25%) proportion of global population to even resulting in human extinction.[6] Nuclear war,[7] or catastrophic climate change[8] are examples of such catastrophic threats. We may also soon face additional global risks from emerging technologies such as synthetic biology and advanced Artificial Intelligence (AI) systems.

The likelihood of many of these risks is highly uncertain and potentially very low. However, we should be careful not to underestimate the chance of any of them occurring. For instance, one probabilistic historical analysis of inadvertent nuclear conflict put the odds at 0.9%.[9] This does not consider the possibility of a hostile first-strike, not of nuclear conflict between other powers. Under a plausible scenario of greenhouse gas (GHG) concentrations reaching 700 parts per million (which would be achieved under the Intergovernmental Panel on Climate Change "Middle of the Road" scenario) there is approximately a 10% likelihood of warming exceeding 6 °C.[10] This does not account for many potential tipping points, and hence the odds could be even higher. Regardless of probability, such risks are of critical importance to society, corporations, and shareholders. Risk is contingent not just on probability, but impact. Rare, impactful events shape the world, including the financial sector.[11] The risks we discuss here can cause tremendous harm, including the dissolution of some of the most powerful industries (or in the case of extinction, all of them). Rather than focusing on setting an arbitrary probabilistic threshold, we suggest focusing on plausible risks: ones which are in line with our background scientific and intellectual knowledge.[12] The risks we will cover here ranging from nuclear war to climate change and lethal autonomous weapons are either already occurring or could plausibly cause a global catastrophe.

Ethically, the threat of large-scale mortality or extinction can be a concern for multiple reasons, ranging from the loss of countless future lives,[13] breaking our obligations to past generations,[14] to simply the harm it would cause to existing, living beings. Most moral value theories can agree that such catastrophes would be wrong, although they differ in their reasons as to why and how bad this would be.[15] While ethics is always contested philosophically, there are good reasons to believe that catastrophe and extinction is largely a point of convergence. This is particularly true for the public common-sense notion of the word rather than the philosophical one.

These global catastrophic hazards overlap with systemic risk, a concept that is widely discussed in corporate finance and business ethics. Systemic risk refers to the ability for individual disruptions to cascade into system-wide failure due to the vulnerabilities and structure of a system.[16] The Global Financial Crisis is the example *par excellence* of financial systemic risk. Systemic risk does not need to be global or catastrophic, but under the right conditions it can be. This is particularly the case when a situation of systemic risk leads to reinforcing, "synchronous failures".[17] While systemic risk is relevant and related, we focus instead on Global Catastrophic Risks.

Beck once reflected that the risk society is born from the unforeseen and unintended consequences of systemic behaviour.[18] He was wrong: global risks are often anticipated, developed and funded by a select few. Companies, and the investors that finance them, are key contributors to anthropogenic global risks. This is true of climate change, nuclear weapons and many emerging dangerous technologies. Since 1988 71% of global GHG industrial emissions can be traced back to 100 companies.[19] Fewer than 30 private companies underpin the maintenance and development of nuclear weapons systems.[20] The development of lethal autonomous weapons take place within an oligopolistic marketplace dominated by tech giants.[21] 39 of the 72 ongoing projects to research and develop high-level machine intelligence (defined as "a general or specific algorithmic system that collectively performs like average adults on cognitive tests that evaluate the cognitive abilities required to perform economically relevant tasks."[22] This is also frequently referred to as "Artificial General Intelligence" (AGI), including in the two surveys mentioned here) are taking place within companies.[23] This marks an increase in private sector projects since 2017.[24] The actions of a small number of investor-owned companies are fuelling future catastrophe.

The implications of global risks for corporate governance and business ethics appear to be chronically understudied. We used the Existential Risk Research Assessment[25] — a machine-learning algorithm which collects global catastrophic risk literature and is updated monthly[26] — to search across a sample (available on request) of 15,000 relevant papers produced by TERRA using the search terms "corporate governance", "business ethics" and "corporate ethics" and find any that directly address corporate governance or business ethics. The one relevant piece we could identify examines the intersection between sustainability discourse and risk management in business ethics,[27] but does not directly deal with Global Catastrophic Risks. Given the significance of catastrophic risks to corporate ethics and profits this is perplexing. We suspect that this lacuna is due to their being little overlap between Global Catastrophic Risk studies and the sub-fields of business ethics and corporate governance. Despite this, some relevant literature exists. There have been some initial efforts to reveal the ownership patterns and financial actor behind global environmental changes such as deforestation of the Amazon Rainforest and boreal forests.[28] But much more is needed to locate the key financial actors involved in creating global "Anthropocene Risks" and determine their obligations and responsibilities.[29] To the best of our knowledge, no study to date has systematically examined the credibility of the underlying arguments for investor campaigns, including for global risks. The Fossil Free movement has received the most academic attention. This has included how it has sowed the seeds of anti-fossil fuel norms,[30] how the movement has operated[31] and its battle for legitimacy with the fossil fuel industry.[32] Despite this coverage the literature has not examined investors' motivations for shifting their investments to address climate disruption, let alone for a broader suite of global risks.

Global risks (excluding burgeoning action on climate change) also appear to be underappreciated in corporate and financial practice. the prevention of global risks has rarely been considered a part of responsible investment. For instance, during 2016 in the UK around £81 billion was invested in funds with "green" or "ethical" principles.[33] None of these funds have enshrined the mitigation of global risk as an explicit principle or objective, and only a small proportion would have represented true "impact" investments.

We address this critical gap in the literature by providing a novel interdisciplinary study and classification of the reasons for institutional investors to manage global risks via their investment strategies, and the tactics they can use to achieve this. Our analysis is valid not only for existing campaigns on climate change and nuclear war, but for campaigns on global risks more broadly. We focus solely on institutional investors and their associated campaigns. This focus covers both institutional investors, and the campaigns by students, activists and others which lead to them taking action.

In Part I we focus on the motivations behind investor campaigns. The analysis proceeds by first discussing what institutional investor campaigns are and what they aim to do. Section 2 examines the multiple different profit-driven and nonprofit rationales for factoring global risk alleviation into investment decisions. Section 3 analyses whether the different reasons for investment redirection hold for all global risks. We find that there are compelling grounds for institutional investors to manage global risk via their investment strategies, built on a range of legal, ethical, economic and political considerations.

Part II examines the tactics of investor campaigns and whether they can be an effective tool for tackling global risks. We proceed by exploring six tactics at the disposal of institutional investors. This is followed by an examination of when investor campaigns are likely to be effective. We then investigate whether investor campaigns could be useful tools for preventing different global risks before concluding with an analysis of ways in which investor campaigns could begin to tackle dual-use technologies.

PART 1 — Institutional Investors' Obligations to Manage Global Risks

1. Background: The purpose of institutional investor campaigns

The activities of corporations have enormous consequences, for good or ill. Several actors have influence over company management and therefore corporate conduct. These include clients, employees, governments, and publics. We focus on one type of entity: investment funds, especially those of institutional investors. Institutional investors are organisations that pool the money of their members and invest on their behalf. Examples include pension funds, sovereign wealth funds, endowment funds (for religious, educational and other non-profit institutions), insurance companies, and banks. They are notable due to their power and motivations, which can be different than those of individual (or "retail") investors or corporate management. Two features that characterise many, but by no means all, institutional investors are their relatively long-term outlook, with many even mandated to preserve the health of their investments in perpetuity, and the breadth and size of their investments, which can come to approximate a representative sample of the economy as a whole.

Over the past four decades "investor campaigns" have become prominent. These involve investors using financing and shares as a way to shift corporate conduct in a more ethical direction. These have included campaigns around the Apartheid regime, tobacco companies, arms companies, nuclear weapons and fossil fuels companies.[34] We consider investor campaigns to include a broad range of approaches, encompassing a suite of tools used to reshape investment patterns, corporate conduct, government policy and consumer behaviour from socially harmful to beneficial activities. These tools range from "shareholder activism" (investors using their voice and influence "from the inside" to directly pressure corporations to change their harmful behaviour) to "divestment" (selling shares in target companies). The commonality across investor campaigns is an explicit aim to brand particular activities or companies as morally wrong, or "sinful", and steer financing away from these activities or companies.

Investor campaigns have experienced mixed results; their effectiveness depends on how their impact is measured. Some activists hope to affect companies' share prices directly through divestment. Yet historical evidence suggests their influence in financially undermining industries such as tobacco, gambling and arms production has been negligible because divestment efforts have tended to focus on public equity, where shares pass from shareholder to shareholder without any exchange of capital with the company itself.[35] Most of the companies in these sectors continue to be profitable and widely invested in. Campaigners have targeted the tobacco industry since the 1980s, yet the industry remains globally profitable and a mainstay among fund managers. While the Fossil Free movement and Don't Bank on the Bomb campaign are spreading rapidly, they have also caused comparatively little financial harm to the companies they are targeting.

However, these campaigns usually do not aim to undermine the share price of destructive industries. Instead, the goal is to eliminate their social license to operate: to make targeted companies into pariahs, to change corporate and consumer behaviour, and, above all, to make them susceptible to more stringent government policy. Partially due to these campaigns, the tobacco industry has now fallen under aggressive excise tax and advertisement regulatory regimes in many countries. Investor campaigns targeted at landmines and cluster munitions were closely involved with the wider political and diplomatic campaigns that led to the international treaties prohibiting these weapons. As Fihn notes for nuclear weapons, "prohibition precedes elimination".[36] Divestment can be regarded as a public shaming tactic, one aimed at changing corporate activities, public discourse and government legislation. It need not directly harm the companies' share prices to be effective. Bergman concluded in a study of the Fossil Free movement that while the direct impacts have been small, the indirect impacts such as changes in public discourse have been significant.[37] While not a silver bullet, social stigmatisation may affect both corporate and government activities.

Investor campaigns such as those that target nuclear weapons and climate change could cover other global risks in the future. Divestment is a strategy that has been employed by social movements since at least the 1980s.[38] Theory and empirical studies suggest that social movement tactics, such as divestment, can spread both between groups within a movement, via intramovement diffusion, and across different movements, via intermovement diffusion.[39] Specific tactics, such as prolonged sit-ins to push for divestment from apartheid South Africa, quickly spread across universities in the US and internationally in the 1980s.[40] Student protests against investments in apartheid South Africa forced IBM, Ford, General Motors, and Exxon Mobil, among others, to withdraw from South Africa. The targets spread to companies associated with arms, tobacco and human rights violations in the 1990s.[41]

Emerging global risks, such as the development of lethal autonomous weapons (LAWs), advanced AI systems and bioengineering technologies, could be the next battlegrounds. Indeed, Pax (the organisers of the *Don't Bank on the Bomb* campaign) are now focusing on preventing an AI arms race.[42] This includes pushing the private sector to "commit to not contribute to the development of lethal autonomous weapons". This raises the question: are investor campaigns justified in targeting global risks? In short, should institutional investors care about global risks?

2. Analysis: Motivations for investor campaigns

This section surveys four existing motivations for investor campaigns related to global risks, and introduces two new ones. The "existing" motivations have been publicly voiced in relation to fossil fuel and nuclear weapon investor campaigns. They include motivations stemming from ethical and legal obligations, including *potential* legal obligations. They rely on either the adoption of new laws in multiple jurisdictions (e.g. codetermination legislation) or a specific legal interpretation of an institution's purpose (the argument for institutional perpetuity). They also include profit-based motivations to reduce volatility and risk. We also introduce two newly proposed motivations: long-term institutional survival and Universal Ownership Theory. These ideas are not novel, but their use as a justification for shaping institutional investment in relation to catastrophic risks is. Institutional investors depend on wider social stability. It is not in their interests to jeopardise or undermine global stability in the long-term. Their interests rely in stability and avoiding Global Catastrophic Risks. These reasons are not discrete and often interrelate. For example, acting ethically (ethical obligations) could improve a company's image (or mend a tarnished one) and aid in profits in the longer-term. Table 1 presents our novel classification of these motivations.

The motivations are generally universal. However, there are some caveats. Organisations' articles of incorporation and bylaws vary greatly, and their obligations rooted in corporate law will vary based on where the company is headquartered, while obligations arising from financial law will also depend on where they are listed or incorporated. The degree of freedom permitted under the Business Judgement Rule will in turn vary according to jurisdiction. Moreover, some organisations are more sensitive to particular rationales. As a general assumption, churches and public bodies such as universities might be more receptive to duty-based reasoning, while corporations might be more persuaded by self-interested arguments that align with their underlying profit motive and fiduciary duties.

Table 1: Motivations for investor campaigns.

Previously Identified or Novel	Rationale	Explanation
Previously identified	Ethical obligation	Investors have a moral obligation not to create significant harm or support others in doing so.
Previously identified	Legal considerations	Corporations currently have legal considerations related to global risks in certain jurisdictions, and could face further obligations in the future. Statutory considerations of codetermination could allow for concerned stakeholders to push for global risk reduction. Benefit corporation laws could be reformed to, or in some cases interpreted to, give institutional investors a fiduciary obligation to reduce global risk as a social goal.
Novel	Perpetuation of the institution	Institutions have a vested interest in ensuring their long-term survival.
Previously identified	Profitability	Investors have an overarching (although not legal) goal of profit maximisation and in some cases divestment can lead to higher returns and lowered volatility, especially for large institutional investors with a long-term view. Not acting ethically damages one's reputation.

Previously identified	Risk avoidance	Divestment can aid long-term profitability by withdrawing from "sin stocks" before regulation is introduced and financial value drops. In some cases, these "sin stocks" risk an absolute loss of value as a sector or subsector is regulated out of existence (or becomes technologically obsolete). Absolute losses wipe out previous gains, making these investments potentially unattractive to cautious long-term investors.
Novel	Universal Ownership Theory	Universal Ownership Theory advances the goals of large institutional investors whose interest is in the long-term performance and stability of the economy as a whole due to their investments across many sectors and asset classes.

2.1 Non-profit-based arguments for investor campaigns

2.1.1 Ethical obligations

Investor campaigns may be driven by an ethical claim: investors should not support, or benefit from, products and services that cause significant social harm. In short, it is wrong to profit from harm. This ethical imperative has been among the most prevalent discourses within investor campaigns. Efforts to change company activities from the inside often appeal to the "better angels" of management. Divestment relies on stigmatising controversial and ethically questionable holdings as "sin stocks". This ethical branding has been used in the divestment campaigns for both nuclear weapons and climate change. The *Don't Bank on the Bomb 2018* report warns financial institutions that support nuclear weapons producers that they will become "increasingly isolated and stigmatised" unless they divest.[43] Similarly, the Fossil Free divestment movement is a battle over hearts and minds.[44] Campaigners aim to discredit and de-legitimise an industry. The campaign has hinged on the simple notion that there are no pensions and no use for degrees on a dead planet. McKibben contends that universities that invest in fossil fuels create a tragic situation in which "educations are being subsidized by investments that guarantee they won't have much of a planet on which to make use of their degree".[45]

Appeals to the ethical principles of investors occur in three key ways. First, they may appeal to institutions' supposed commitments to generalised ethical principles, which can reflect a diversity of ethical traditions and schools of thought.[46] Second, they may demand consistency between an investment and the stated principles of the investor or shareholders. This is particularly important for institutions with higher ideals, such as a university, government or religious institution. Deviation from these principles, or the standards of wider society, can grievously injure the reputation of a company or investor. Third, they may influence the individuals involved in the institutions. Executives at companies and institutional investors are not purely economic agents; they consider their reputations and identities when acting.[47]

Additionally, institutional investors have a duty to their stakeholders not to put them at undue risk, including from global catastrophes that they have the power to influence. By tacitly supporting corporations that are contributing to global risks, these institutional investors are in turn potentially placing their stakeholders in harm's way. Investors' responsibilities to their stakeholders should not be limited to purely financial matters where other pressing interests of theirs are at stake.

Ethical obligations could also stem from theories of corporate governance. Stewardship theory sees management and leaders as having a responsibility to guide and protect the long-term performance of a firm.[48] Others have sharpened this to "ethical stewardship": the "honoring of duties owed to employees, stakeholders, and society in the pursuit of long-term wealth creation".[49] Such an approach to leadership lends itself to protecting against catastrophic risks. Preventing global calamity is clearly in the interests of both society and long-term wealth creation, and as highlighted in Section 3.2 is compatible with, if not supportive of, firm performance. An ethical steward would ensure that their employees, society and stakeholders do not face undue, dire risks.

Note, that this is a description of how ethical obligations are articulated, and why they are reasonable. They are not an argument that companies will act ethically. Indeed, there is ample evidence that unethical conduct is brazen and rife in many industries including pharmaceuticals, arms production, and finance.[50] This is not a naive plea that institutional investors will comply with ethical demands. Rather, that these ethical obligations exist, are sensible, and are one of many pressures that can change the behaviour of institutional investors and companies.

Following these ethical arguments, which can only be enforced through individual consciences and the court of public opinion, the next subsection will explore the potential of arguments that can be enforced in courts of law.

2.1.2 Legal considerations

Institutional investors are currently under some legal obligations in relation to global risks. There are several possible pathways for them to be considered under other legal obligations.

(a) Considerations relating to domestic and international legal obligations

Institutional investors fall under various existing legal obligations in relation to global risks. These often depend on legal structure and the jurisdiction in which they are incorporated or listed. For example, fiduciary duty for pension trustees is increasingly being interpreted as incorporating the duty to include climate risk into investment analyses. In the UK, the Bank of England specifies how financial institutions are obliged to account for climate change risks.[51] ClientEarth recently reported four major UK companies to the UK regulator, the Financial Reporting Council, for failing to address climate change risks in their shareholder reports.[52] Similar legal hurdles are arising in the world of nuclear weapons. A company operating in a country that has ratified the Treaty on the Prohibition of Nuclear Weapons will be bound by the provisions against contributing to the development, testing, production, stockpiling, stationing, and transfer of nuclear weapons. Legal obligations exist in both international and domestic law.

(b) Considerations vis-à-vis employees

Corporations and other institutions have existing legal (*de lege lata*) obligations to their stakeholders that could underpin claims to change corporate behaviour or divest. Most larger corporations have hard statutory obligations towards their employees and other stakeholders beyond labour contracts, workplace safety and future pension entitlements.

More specifically, laws of codetermination provide workers with a legal right to participate in the management of companies in which they are employed. Such a model is already practised in numerous jurisdictions such as Germany, Austria, Sweden, France, Denmark and the Netherlands, with the German Codetermination Act (*Gesetz über die Mitbestimmung der Arbeitnehmer*) of 1976, based on an earlier law from as early as 1951, representing a prominent statutory example. The purpose of the codetermination model is to represent the interests of workers alongside the predominating interest of shareholders. Current employees normally have a vested interest in avoiding global risks, both in preserving their own health and well-being as well as that of future generations. In the future, long term-oriented employees might begin to use co-determination to oblige their institutions to move away from risky activities and towards socially beneficial ones. This has yet to occur, but is a promising legal avenue. In this case, co-determination laws are an enabling factor, but action still rests on employee motivation and power relationships.

(*c*) *Considerations relating to fiduciary duties for profit and return*

A common objection to divestment and company or sector exclusions is that institutional investors usually cannot legally divest or exclude holdings for moral reasons as this would violate trustees' duties to beneficiaries. For instance, the Business Judgement Rule states that executives have a fiduciary duty to act in good faith, loyally, with due care (Cede & Co. v. Technicolor, Inc., 634 A.2d 345, 361).[53] Similar rules exist in five European states, namely Germany, Romania, Croatia, Greece and Portugal,[54] and their obligation to maximise profit for shareholders. Similarly, institutional investors sometimes claim that they cannot engage in campaigns due to fiduciary duties to their trustees to maximise returns and maintain a sufficiently diversified portfolio.

These contentions are all highly contested. Legal experts and scholars have laid out a compelling case in multiple jurisdictions that not only is divestment from certain companies or sectors permitted[55] but investing for impact may be legally required. This is because investment returns — and benefits to beneficiaries — rely on the health of the overall economy.[56] Some scholars have suggested that the singular duty to maximise returns is an ideological myth and that fiduciaries instead must meet several different fiduciary obligations.[57] Importantly, even if this were true there is little evidence that divestment injures profits for companies or returns for investors (see Section 2.2). There are similarly no strong grounds that divestment will injure the diversity of a portfolio substantially enough to violate duties to trustees. This is clear in the long history of investors that have divested from numerous areas (such as tobacco, arms, and gambling) without substantially reducing portfolio diversification.

2.1.3 Protecting investments in perpetuity

Many institutions were founded with the implicit or explicit aim to continue in perpetuity. This includes many schools, universities, and colleges,[58] religious institutions,[59] NGOs, and some charitable endowments, referred to in the UK as "permanent endowments".[60] Their founders clearly envisioned that they would continue forever. Others, such as sovereign wealth funds and pension funds, are predicated at least on long-term, if not perpetual, existence. The Norwegian Oil Fund's mission statement is "to safeguard and build financial wealth for future generations".[61] Similarly, much of international law is based on the assumption that states are perpetual.

These institutional investors therefore have a special obligation to ensure the survival of their institution and guard against risks that might destroy it. Global risks that would produce a substantial disruption to the global economy could lead to the bankruptcy or destruction of many corporations and investment vehicles. Indeed, these global risks are some of the few events that could lead to such an outcome for significant swaths of an investor's portfolio. Thus, investors have an obligation to ensure that they are not contributing to them occurring. The investments of these institutions should be compatible with a vision of themselves as a long-term or perpetual institution. To the best of our knowledge, this is a novel and overlooked rationale that has not been prominently used in existing investor campaigns or discussions.

This is not an argument that commitments to perpetuity will *de facto* lead to a sober consideration of global risks. For instance, the majority of capital inflow for the Norwegian Oil Fund comes from oil and gas extraction. Its commitment to safeguard future generations is at odds with its contributions to climate change. This is an argument that pledges to exist in perpetuity are one (often underexploited) reason that these actors *should* be acting to address global catastrophic risks.

2.2 Profit-based arguments for investor campaigns

2.2.1 Profiting from investor campaigns

There is tentative evidence that socially responsible investment portfolios may be just as profitable as irresponsible ones. This has been substantially explored in the case of portfolios excluding fossil fuels. Trinks et al. compared portfolios with and without fossil fuel stocks over the period 1927–2016 and found that a fossil-free portfolio's performance was similar to that of a portfolio including fossil fuel stocks.[62] Grantham conducted a comparable study looking at nine major sectors in the stock market and the US-listed companies included in the S&P 500 Index over the past three decades.[63] He found that excluding any single sector made no significant difference in portfolio returns. Excluding energy actually increased returns by three basis points. This finding is supported by several earlier studies showing no or little impact of fossil fuel screening on portfolio performance;[64] contrary to some claims, divestment does not appear to weaken financial performance. In fact, one study found that a fossil-free portfolio (S&P, excluding fossil fuel companies) outperformed the S&P 500 Index and a fossil-fuel orientated portfolio within the S&P 500 over an eight-year period between 2010 and 2018.[65] Finally, the Cambridge divestment report summarised all of the above studies and all of the other peer-reviewed analyses available, concluding that although it is possible to time fossil fuel divestment well or poorly because fossil fuels have outperformed or underperformed the market at different points, judging from a review of studies covering 118 years of data there was little overall difference between fossil-free and conventional portfolios.

Evidence on the performance of socially responsible investments is similarly mixed. Dimson, Karakaş and Li find that after shareholder engagement, particularly on environmental and social issues, companies experience improved accounting performance and governance as well as increased institutional ownership.[66] Yu looked at the monthly risk-adjusted returns for 321 funds over 1999–2009 for both social-screened and conventional mutual funds.[67] He found that funds in the ethical governance and social categories outperformed their conventional peers. However, there is a debate in the literature as to whether socially responsible portfolios outperform conventional ones.[68] For example, some studies suggest that investors who have maintained shareholdings in tobacco may have benefited disproportionately relative to those who divested.[69] In general it is still debatable whether socially responsible investment funds outperform, underperform, or match market performance.

There is a connection between the profit-motive and ethical obligations. That is, breaking ethical norms can injure a firm's reputation. As mentioned earlier, this loss of reputation can affect an investor's or company's bottom line by enabling regulation, or by leading to boycotts and consumer backlash.

2.2.2 Avoiding legal and reputational risk

If a particular line of business or "sin stocks" will likely become a pariah, or become the object of stringent regulations, then shifting corporate behaviour or withdrawing early can avoid losses. Institutional investors can evade risk by encouraging change or divesting.

Selling stocks before a tidal market shift can be lucrative. Several investors made small fortunes by foreseeing the 2008 subprime mortgage crisis and shorting the market.[70] These are cases of investors profiting from a market collapse rather than widespread divestment, but the fundamental premise is the same: it is possible to generate profit and/or avoid risk by diverting funds before a hidden systemic risk is realised by the wider market. The idea is simply to escape the bubble before it bursts.

Such bubble thinking has been prevalent in the Fossil Free movement. It was influenced early on by Bill McKibben's Rolling Stone article 'The Terrifying New Math of Climate Change'.[71] His popular piece drew from a report by Carbon Tracker on 'Unburnable Carbon'.[72] Unburnable carbon refers to fossil fuel stocks that need to remain buried and unused to limit temperature rise to 2°C. One study by McGlade and Ekins suggests that unburnable carbon could represent known reserves as high as 96% of coal, 54% of conventional oil (100% of unconventional) and 69% of conventional (82% of unconventional) gas.[73] Thus the Fossil Free movement argues that, due to this unburnable carbon, investments in the fossil fuel industry are financially reckless.[74] The stock price of these companies is built on inflated valuations of unburnable carbon and stranded assets. This asset value will need to be prematurely retired and written off to meet the goals of the 2015 Paris Agreement. Such losses could be significant. Mercure et al. estimate that the losses from stranded assets may amount to a discounted global wealth forfeiture of US $1–4 trillion.[75]

We argue that the same approach can be applied to "sin stocks" beyond fossil fuels. Trying to change the activities of (or create an early exit from) controversial companies is rational if there are good reasons to believe that government intervention, technological advances, or a divestment contagion will permanently drop the price. This could occur with fossil fuels, nuclear weapons, lethal autonomous weapons and advanced biotechnological and AI systems. Companies in these areas could be in a bubble of underpriced risk. Divestment or dropping particular lines of business entails escaping the bubble before the risk is addressed by government or the judiciary. Which branch of power, the legislature, the judiciary and the executive, creates the legal risk will vary by jurisdiction and case. In Japan the bureaucracy is empowered to change the financial industry. In the US the power rests primarily in the courts.

However, there is some scepticism as to whether one can divest oneself out of a bubble, carbon or otherwise. First, the timing is notoriously difficult to get right. This is evidenced by the rise of passive investing at the expense of active investing (where stock-picking, and its timing, are of paramount importance). Second, some evidence suggests that climate risk may be largely "unhedgeable". One simulation of a variety of types of diversified portfolios found that it was possible to hedge against less than half of climate risk.[76] The nature of Global Catastrophic Risks or existential risks is that they affect the planet as a whole, and therefore cannot be avoided through clever or prescient stock-picking. Whether risk avoidance will work thus depends on the extent and type of market disruption.

Legal risk can also stem from litigation. The advent of attribution studies could provide the foundation for a wave of lawsuits against the fossil fuel industry. Studies that attribute extreme weather events to climate change are already commonplace and improving.[77] Given the highly concentrated nature of the fossil fuel industry and supply side emissions, the question of legal liability becomes apparent. For example, Ekwurzel et al. have attempted to attribute responsibility for particular climate effects to particular companies.[78] Already suits have been brought against major fossil fuel companies for damages caused by climate change in both New York City and California.[79] Greenpeace has threatened similar action in the Netherlands against Royal Dutch Shell plc.[80] In 2015 the Urgenda Foundation and 900 Dutch citizens won a case against the government of the Netherlands, forcing it to improve its emissions reductions efforts to fulfil its duty of care in protecting Dutch citizens from climate change (Urgenda Foundation v. The Netherlands [2015] HAZA C/09/00456689 (June 24, 2015)). The case was appealed, but upheld in 2018 (Aff'd (Oct. 9, 2018) (District Court of The Hague, and The Hague Court of Appeal (on appeal)).

The outcomes of such cases will be contingent on the jurisdiction and circumstances. The suit in New York City has already been dismissed, although this was because the judge deemed it to be under the purview of the Federal Court.[81] Such cases can be expected to increase over time as the impacts of climate change worsen. Even if unsuccessful they can and often do cause reputational harm to companies and increase the chance of other investors withdrawing or diverting investment. Cases have also begun to be brought against investors, such as in Australia and the UK in recent years. Furthermore, if a company loses a lawsuit and fines are levied, the costs be passed on to investors via cuts to dividends and decreases in share prices.

There is a second risk: reputational. A company or investor's image is increasingly important to their prospects. Alphabet Inc. was quick to withdraw from the US military contract of "Project Maven" once they faced a boycott from their employees. The Pentagon project aimed to develop algorithms to differentiate between objects and people based on big data from military drones. The threat of losing highly skilled, conscientious employees outweighed the money on offer from the Pentagon. One key historic example was a $5 million contract to produce napalm for use in the Vietnam War, "which most likely cost Dow Chemical billions of dollars" in "damaged reputations, recruiting problems and customer boycotts".[82] This ability to attract and retain consumers and employees has been included in "intangibles" accounting. Corporate executives know that financial incentives alone are not enough to recruit and retain skilled employees.[83] One review of the literature suggests that a company's image plays a role in prospective employees' decision-making processes, that they prefer to work for corporations whose values overlap with their own, and that this affects long-term employee retention.[84] Such considerations are particularly important for top talent.[85] Being perceived as a funder of global risks and a questionable corporate citizen is a financial and recruitment liability.

In summary, the combination of legal and reputational risks makes aligning investments to minimise global risks a prudent choice for institutional investors.

2.2.3 Universal ownership theory

"Universal owners" are long-term asset owners such as pension funds or sovereign wealth funds that are invested in a broad (and more or less representative) swath of the economy.[86] These universal owners have highly diversified portfolios that span different asset classes encompassing public and private debt and equity, physical assets, and more. Universal ownership theory supposes that an investor who is invested across a broad swath of the economy will have an interest in its overall health. Externalities from one sector could affect returns in another, especially in the long-term, which necessitates a more holistic approach.[87] For example, high emissions in the fossil fuel sector could have negative consequences for returns in other asset classes. This includes physical effects on infrastructure, health effects due to pollution in cities and decreases in worker productivity. These factors could easily affect the long-term returns of other parts of a universal owner's portfolio. If harm from one company or sector affects other companies or sectors within a universal owner's portfolio, the owner has a strong interest in reducing that harm. For a universal owner, it would make sense to work to reduce the harms produced in the fossil fuel sector to protect the rest of the portfolio.

For the universal owner, who is by definition a long-term investor with a system-wide view, externalised outcomes are counterproductive. Trade-offs between and among companies and sectors need to be accounted for. As universal owners have an interest in the overall health of the economy, it is self-defeating for them to allow the companies it is invested in to contribute to global risks. Instead, they would be better advised to internalise the externality, and use their power through an investor campaign to change companies' behaviour.[88] Universal ownership theory thus provides strong theoretical grounds for a large subset of institutional investors to shift their finances away from contributing to global risk.

3. Are the motivations applicable to all Global Catastrophic Risks?

All of the outlined rationales are compelling, but not all are generalisable across all institutional investors or global risks. The two novel motivations we list (long-term institutional survival and universal ownership theory) are only applicable to particular types of institutional investor. That is, either those with an enshrined goal of perpetuity for the former, or a heavily diversified asset owner portfolio across a representative proportion of the economy for the latter. Arguments to reduce risk or maximise profit vary by issue. Divestment appears to have been a financially prudent move for fossil fuels, but not for tobacco. Similarly, legal, regulatory and financial risks will vary by country and market.

Ethical arguments appear to be universal across actors and risks. All global risks share the common "traits" of destroying vast amounts of future material and immaterial value. The potential for large-scale loss of human life makes investments in activities that contribute to these risks unethical. It is immoral to fund or profit from the endangerment of future generations. That ethos holds true regardless of the threat. Similarly, duties for institutional perpetuity do not depend on the hazard involved. There are no shareholders, churches or universities in a collapsed civilisation. These duties do not hinge on the nature of the specific risk, but rather its potential for causing future damage. Overall, institutional investors appear to have some core common ethical duties to reduce global risk, but whether they are legally bound to, or will profit from it, differs.

The different motivations also vary in terms of their effectiveness. Arguments for ethical action, the avoidance of reputational damage and regulatory risk have been persuasive in the case of both climate change and nuclear weapons. For example, the Norwegian sovereign wealth fund, Government Pension Fund Global, has excluded 16 companies involved in the production of nuclear weapons due to their potential to violate humanitarian principles.[89] The Rockefeller Brothers Fund (RBF) cited both the moral tension between supporting the fight against climate change and investing in fossil fuels, as well as the fiscal responsibility to avoid stranded assets, as the basis for its decision to divest.[90] There is abundant evidence on the efficacy of ethical shaming and stigmatisation tactics.[91] This is evident in the proliferation of corporate watchdogs such as the Multinational Monitor, Corporate Watch and Global Exchange.

Redirecting institutional investment to maximise profit, avoid legal risk or meet legal obligations are motivations with more uncertain results. As noted, most legal cases related to fossil fuel divestment have been unsuccessful, while there is little evidence to suggest that diverting investments from globally risky activities will be profitable. Put simply, more time and evidence are needed to assess the profitability of a global risk-averse approach. The case for reducing global risks through investments to protect institutional perpetuity or broader economic stability (universal ownership theory) is even more unclear. These are novel motivations that have rarely been publicly advocated for. Some reasons are more compelling than others, yet none is entirely without grounds. The profit and non-profit basis for a global risk-averse investment strategy appears to be varied and sound.

PART 2 — Institutional Investors' Tactics for Managing Global Risks

4. Analysis: Six tactics for investment campaigns

Investor campaigns typically employ six tactics: contest, protest, request, divest, re-invest and acquest. Each of these approaches has different strengths and weaknesses. In the following overview presented in Table 2, we describe these tactics and their aims. We focus on tactics relevant to institutional investors and financial redirection. These tactics run across the spectrum from least (contest) to most (acquest) confrontational. There is a long-standing debate on the relative effectiveness of more confrontational tactics: from shareholder activism to divestment.[92] However, it seems that a range of tactics operating in tandem can work. The "inside game" can support the "outside game" and vice versa. Examples include tobacco, apartheid, landmines, cluster munitions and corporate social responsibility (CSR) in general).[93]

Table 2: The six tactics of investment campaigns.

Tactic	Description	Aim	Conditions
Contest	Refers to shareholder activism by those with voting shares who aim to directly influence decision-making, involving a (group of) shareholder(s) with non-negligible holdings, using internal mechanisms to steer a company away from irresponsible, unethical or risky actions; includes both voting and co-filing resolutions and encompasses attempts to introduce independent monitoring and verification schemes to reinforce the desired behaviour.	To shift a company from within through internal governance processes.	The company must be capable of changing its behaviour, i.e. its business model cannot entirely dependent on the activity. There should be non-negligible holdings of voting shares, and the company's terms of incorporation must allow for this form of influence.

Protest	Refers to shareholder activism by those with shareholdings who aim to lobby a company. For instance, by directly contacting senior executives or using Annual General Meetings (AGMs) for publicity and message amplification, involving a shareholder using their position to raise critiques and urge company transformation in annual general meetings.[94] May be done by purchasing a single share. Protest also includes shareholder litigation. That is primarily derivative actions and other corporate-law-based instruments to hold executives responsible and liable by way of a suit in court against decisions that harm the (long-term) interests of shareholders. Both shareholder suits against the corporation and against individual executives are possible, both on behalf of an individual shareholder and on behalf of the corporation itself (i.e. derivative action). Even if such suits are not likely to be successful, their mere filing can attract public attention to management decisions that increase global risks.	To cause a change to company behaviour or a loss of social licence from within by critiquing harmful activities in boardroom meetings and from outside by releasing this to the media or using shareholder litigation. Shareholder litigation can also shift company activities through the threat of regulatory, financial or reputational damage to the company or particular individuals.	The company must be possible to change; negligible holdings may suffice. In the case of litigation the feasibility of a case will vary by jurisdiction and circumstances.

Request	Refers to shareholders and institutional investors directly appealing to government to introduce policies that reshape financial flows, impose standards or enact selective purchasing laws based on promoting their interests as investors. For example, at the 2018 Katowice climate summit, 420 investors managing over USD $32 trillion in investments called for stringent emissions reductions policies and improved climate-related financial reporting measures.[95]	To push governments towards policies causing investors to change asset allocation and businesses to change behaviour.	An industry's political power must be weak enough to allow for effective regulation, or the power of investors must be sufficient to overcome industry-imposed barriers to this.
Divest	Refers to shareholders freezing, reducing, or fully disposing of their holdings; this entails both the action of divesting by shareholders, as well external campaigns that urge divestment.	To avoid risks, maximise returns, socially align one's portfolio or cause the loss of social license. Within public equity the effect is indirect in that it is mainly due to public pressure; in other asset classes it can actually shift capital, and therefore can be more directly impactful.	The company must be unable or unwilling to change; any prior holdings suffice.

Reinvest	Refers to positive investments in direct[96] alternatives that are risk-reducing at a system-wide level. E.g. for climate change this would be renewable energy; as noted earlier, some positive alternatives can offer equivalent returns and reduced risk.	To redirect capital from socially harmful or risky activities towards alternatives that directly alleviate these risks or damages.	The previous company could not or would not change; or as standard practice.
Acquest	Buying up a company so that it comes under the direct control of an institutional investor. This can be done for the entirety of a company, or potentially just a portion of its shares, product lines or infrastructure.	This is done with the intention of changing the company's business model away from socially harmful activities, or shutting it down.	The firm must be incapable of changing its behaviour and other tactics are not working; the investor must have financing to buy out the company. The legality of this tactic will vary by jurisdiction.

Our list of six tactics can encompass numerous other approaches not explicitly listed here. For example, reinvestment could also involve the creation of new alternative financial institutions that invest in positive rather than risky activities.[97]

An overview of how these different tactics interact is provided in Figure 1. It shows that divestment and positive investment are two sides of the same coin. Contest and protest encourage companies to change their activities, while divestment frees up resources to be redirected towards positive alternatives (reinvestment). It also provides political pressure that can be translated into behavioural or policy change. Insofar as the social licence to operate has been undermined, request tactics will be more effective at enacting more forceful regulation.

Tactics can be combined throughout a campaign, moving from one to another in reaction to corporate lack of change. For example, the financial services company Legal and General Group Investment Management warned of its intention to divest from non-compliant companies two years prior to taking action. This gave targeted companies the opportunity to change their actions and align with the new standards. Thus, the move followed "protest" with "divest". Further research could usefully address whether this was more effective than either tactic alone.

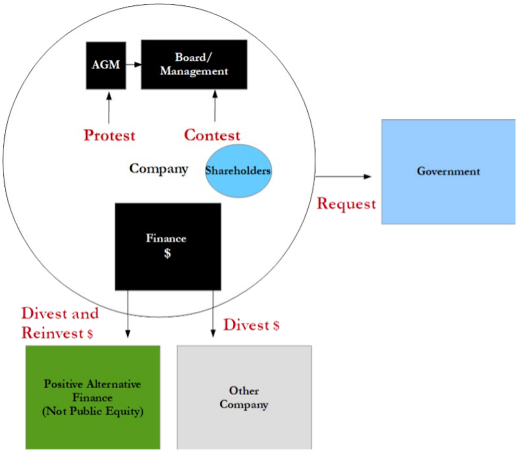


Fig. 1: An overview of investor campaign tactics.

There is a long-standing debate on the effectiveness of engagement compared to confrontation.[98] We believe our typology offers a new angle on the "shareholder activism" versus "divestment" debate, by suggesting which tactics might be appropriate and successful in particular situations. We suggest that shareholder activism (i.e. contest and protest) is appropriate if such activism can reasonably be expected to change a company's harmful activities within a reasonable time period (such as five years to a decade, depending on the urgency of the threat). However, if such activism cannot reasonably be expected to change the company's activities over this time period, then "request", "divest", "reinvest" and even "acquest" are appropriate.

5. Timing of tactics: When to escalate

Whether or not it is reasonable to expect shareholder activism by a particular investor campaign to change a particular company's activities is a difficult question. It will ultimately always be a subjective judgment call. However, some factors can inform this question of whether the company is able and willing to change. These include the centrality of the activity to the business model of the company, and the past behaviour of the company. These factors were identified in 11 interviews with people involved in investor campaigns, all conducted under the condition of anonymity.[99] There are four main conditions to determine when to initiate or escalate shareholder activism. These are: ability to change, willingness to change, timing, and susceptibility to activism.

5.1 Ability to change

If a company's business model is built on activities that create global risk, then there is little hope for it to change. Sometimes the risk-increasing line of business is not central, however. For example, many arms companies can do without nuclear weapons and LAWS contracts; these sources of income are not central to their business model. Large technology companies can very easily do without LAWS contracts. They already have significant profit margins and market capitalisation, and are not reliant on military clients. Some small or medium-sized fossil fuel-related companies, such as the power company Vattenfall AB, might be able to change their business model to become sustainable energy companies. In contrast, driving global warming appears to be irrevocably central to the model of most major fossil fuel companies. The sheer sunk costs in expertise, infrastructure, and assets developed to find and extract hydrocarbons is overwhelming.

5.2 Willingness to change

Is the company engaging with activism in good faith, or merely greenwashing? Both ability to change and past behaviour can be useful guides to this question. For example, fossil fuel companies have a long track record of misrepresenting science and funding "merchants of doubt".[100] They are currently spending billions lobbying the EU and US governments.[101] The key metrics for willingness to change all relate to the question of where the company's new financing is going: the percentage of capital expenditure, research and development spending, and acquisitions (and disposals) spending dedicated to changing course away from Global Catastrophic Risks;[102] thus far none of the oil and gas majors are on track to shift their business models according to these metrics. Combined with their structural inability to change, this provides a strong case that the fossil fuel industry is also simply unwilling to change.

5.3 Timing

How long has there been a shareholder activism campaign and are there signs it is working? There is no clear rule or threshold here. It will partly depend on indicators of progress and how urgent the risk being addressed is. If an insider campaign has been ongoing for multiple years without any clear victories, it is likely time to escalate.

5.4 Susceptibility to activism

How likely is a company to be influenced by shareholder activism? Consumer-facing companies may be more susceptible, especially if their consumers are ethically conscious shoppers. Companies that are more reliant on highly skilled, well-organised workers in short supply may be more susceptible as well.[103] Companies with a corporate culture of, and management incentive structures for, being agile, disruptive and high growth rather than complacent and defensive may be more susceptible. Ownership and management structure are also important. Companies with an individual majority owner, such as founder-dominated technology companies, can move fast if that individual is persuaded.

These are partly empirical questions, but the final decisions are ultimately subjective judgements. There are also practical considerations, such as the potential speed of divestment, impending regulation and public opportunities to garner attention. If a company is unwilling to change, and simply incapable of doing so, then it would be prudent to turn to divestment and reinvestment. If possible, it may even be wise to consider the tactic of acquest.

6. Effectiveness: When do investor campaigns work?

Investor campaigns of the kind described above will not always be effective. As we will substantiate below, evidence from practice and the literature suggests that investor campaigns work best under the following four conditions: the presence of clearly identifiable, intentional moral villains (actors that plausibly contribute to global risk) to target, existing alternative models of best practice, the presence of a well-organised and well-resourced campaign, and private investors who have substantial leverage over the actors. We term this the "Villain-Hero-Campaign-Leverage" (VHCL) framework.

6.1 Intentional "villains"

Political messages inevitably need to be less nuanced than those reflecting the full complexity of the real world. Experience has shown that the presence and visibility of moral villains acting with intention makes ethical arguments salient and political tactics psychologically effective. The use of a simplified narrative and opposing heroes and villains is a central element of investor campaigns. Benford and Hunt contend that "social movements can be described as dramas in which protagonists and antagonists compete to affect audiences".[104] To be effective, framing by movements needs both "heroes" and "villains".[105] For example, a compelling narrative framing has been shown to be more effective in the communication of climate change.[106] The necessity of having a bad actor with intention for a campaign to target is evident in the success of previous campaigns. For example, the Fossil Free divestment movement directly labels the fossil fuel industry as the enemy. The Don't Bank on the Bomb campaign focuses on arms companies, and highlights the emotionally salient *hibaku-sha* victims of Hiroshima and Nagasaki. The apartheid divestment campaign had a ready-made villain in the discriminatory South African government, while anti-smoking activists focus on shaming Big Tobacco. The ability to portray a salient villain with intention allows an investment campaign to construct a persuasive narrative.

6.2 Reinvestment heroes

A successful framing should also include "heroes": actors that are working for the societal good to mitigate global risks. Empirical evidence underlines the unique importance of a hero. Jones used an internet experiment with 1,500 US citizens to explore how different considerations shape individual risk and policy preferences for climate change.[107] He found that narrative framing, particularly the presence of a compelling hero, was the greatest shaper of individual views on climate risks and policy. Investment campaigns ideally pinpoint a clear alternative that mitigates the targeted risk or harm. To change corporate behaviour through the tactics of contest and protest, it is best to have a clear "ask" and proposed model of best practice. The tactic of reinvesting needs a "hero" for reinvestment; having such a hero can improve the efficacy of the tactic of divestment. In the case of climate change, investments in fossil fuels can be reallocated to renewable energy projects (as long as these are primary market investments — not public equity). This could happen within a given energy company, but mostly the Fossil Free movement has employed a frame of shifting money from the "bad actors" (the fossil fuel industry) to good ones (the renewable energy industry). These heroes are the final destination of funds mobilised for positive investment. However, this is not a necessity. Nuclear weapons do not have a clear, profitable mitigating alternative, yet ICAN's Don't Bank on the Bomb campaign has operated successfully regardless.[108] Nuclear weapons producers have been excluded from a USD $1,537 billion asset pool as both the Norwegian Government Pension Fund, ABP (the world's fifth-largest pension fund) and 22 other institutions have enacted comprehensive bans on investing in nuclear weapons. This progress has been underpinned by both the continued activism of ICAN, as well as the symbolic power of the 2017 Treaty on the Prohibition of Nuclear Weapons.[109]

6.3 A well-organised and -resourced campaign

Effective action by institutional investors tends to be underpinned by a well-organised and -resourced campaign.[110] High-profile shareholder activism organisations include As You Sow, Majority Action, and the Interfaith Center on Corporate Responsibility. For fossil fuel divestment the most well-known campaign is the Fossil Free movement centred around 350.org, while for nuclear disarmament it is the Don't Bank on the Bomb campaign. Such campaigns can help to publicly and forcefully express the rationale for divestment campaigns,[111] coordinate a variety of disparate investors and lower the start-up costs (including in terms of risk and transaction costs) for first-movers.[112] These first-movers help to promulgate a new standard across the community that becomes legitimised over time.

6.4 Investor leverage

These six tactics are most useful when deciding about investments in the primary market — private equity, venture capital, bond issuances, infrastructure, private real estate and public equity investments at the Initial Public Offering (IPO) stage.[113] Within the primary market shareholders have the greatest influence over company behaviour, and divestment decisions are more likely to have an impact on the liquidity, cost and availability of new capital, and profits of a corporation or industry. These tactics can also be applied to some extent to publicly traded companies, where different divestment and engagement strategies can still play important public relations roles.

7. Can investment campaign tactics work for global risk prevention?

In this section we examine how the six tactics of investor campaigns can be applied to four other global risks (in addition to climate change and nuclear war) stemming from biotechnology, LAWS, advanced AI systems and asteroid strikes.[114] Biotechnology, LAWS and advanced AI systems are capable of increasing global risks in the coming decades. Asteroid strikes are a naturally occurring risk that has never been viewed from the lens of institutional investment but that we include as representative of natural risks in general.

We review each of these risks according to the four conditions outlined in our framework above: the ability to identify and portray a clear villain with intention, direct (heroic) alternatives, the presence of a campaign, and investor leverage over the problem. Our findings are summarised in Table 3. This includes a comparison against the two global risks that are currently the target of divestment campaigns: climate change and nuclear weapons.

Table 3: Estimating the suitability of investment campaign tactics to mitigate global risks.[115]

Hazard	Presence of an Investment Campaign	An Easily Identified Salient Villain	Heroes for Reinvestment	Investor Leverage
Climate change	Yes	High	High	High
Nuclear war	Yes	High	Low	Medium
Biotechnology	No	Mixed	Yes	Medium
LAWs	No	Yes	No	High
Advanced AI systems	No	Mixed	Mixed	Medium
Natural risks, e.g. asteroid strike	No	Low	Low-medium	Low

7.1 Biotechnology

Biotechnological risks are unprecedented anthropogenic risks. They range from intentionally created weapons through to accidentally birthed pathogens. One risk from emerging technologies is an engineered pandemic. Bioengineering could provide the means for creating a pathogen that is more virulent and deadly than any found in nature. Such a disease could be released through "error or terror" and cause mass fatalities and economic damage.[116] The threat is not speculative. One postdoctoral researcher was almost single-handedly able to produce a complete synthesis of a horsepox virus — similar to smallpox, which killed 300 million people in the 20th Century — in only six months.[117] There is a small but significant chance that these risks could be global and catastrophic.[118]

While representing only a small part of the overall risk, there are some identifiable "villains" in this field. Over the course of the 20th century, 23 states "had, probably had, or possibly had" a biological weapons program, although only those of the Soviet Union and the United States developed significant[119] capabilities".[120] Non-state groups such as Aum Shinrikyo or Al-Qaeda have also sought to develop biological weapons.[121] These terrorist groups, and the Soviets who violated the Biological Weapons Convention for years after signing it, can fairly be called "villains". In many other cases, risks were created despite good intentions and would only create harm due to an accidental release. For example, in 2011 several research groups produced a strain of H1N1 avian flu that was potentially transmissible between humans.[122] These "gain-of-function" experiments to create potential pandemic pathogens (PPP) have attracted controversy.[123] These researchers are not "moral villains". Such research is often aimed at reducing biological threats by better understanding their nature. While there are state and non-state actors looking to use biotechnology to create weapons for the sole purpose of inflicting damage, other threats are born accidentally from well-meaning research. Given this, we rate the presence of a plausible moral "villain" as mixed.

The leverage of institutional investors is often low due to the prevalence of government-funded weapons programs, academic or government-led research and terrorist activities. However, companies appear to be playing an increasing role in biotechnology. For example, the International Gene Synthesis Consortium (IGSC) is an industry-led group of gene synthesis companies and organisations formed to design and apply a common protocol to screen both the sequences of synthetic gene orders and the customers who place them. They represent approximately 80% of commercial gene synthesis capacity worldwide. The centrality of such industry research in potentially risky areas could make investor campaigns an increasingly useful tool in the future.[124] Given this we score the current state of investor leverage as "mixed".

There are clear risk-mitigating alternatives (or "heroes for reinvestment") in the private sector. These include companies supporting better health surveillance, pandemic preparedness initiatives and vaccination production facilities. There is not yet a campaign on this issue. Thus, we rate biotechnology as "mixed" for the presence of a villain with intention, "yes" for the presence of a clear alternative, "no" for a current campaign, "high" for tangibility of assets and "medium" for investor leverage. If companies become a more important player in the future and are engaged in risky behaviour, then an investor campaign might well be appropriate and successful.

7.2 Lethal Autonomous Weapons (LAWs)

LAWs are weapons systems capable of autonomously identifying, selecting, and killing targets without meaningful human intervention or control.[125] They are a near-term technological development and could pose a threat within the next decade. Robotic systems with limited autonomy already exist and are regularly deployed in combat, including the Phalanx Close-In Weapon System and anti-tank and personnel mines.[126] In 2017, 49 deployed systems (from an analysis of 154 systems with automated targeting) could detect possible targets and attack them without human intervention.[127] An open letter signed by over 3,700 AI and robotics researchers and over 20,000 others claimed that cheaply mass-produced LAWs would be the "Kalashnikovs of tomorrow".[128] That is, ubiquitous on battlefields and easily accessed by terrorist groups. The development, stockpiling and/or use of LAWS could change the cost and speed of wars, and therefore destabilise the global order and/or spark escalating arms races. Moreover, they could empower dictatorships with a new form of brutal control and terrorists with a potent, cheap, and mobile weapon.[129] Finally, their vulnerability to unexpected interactions could lead to "flash war", analogous to "flash crashes" in the stock market.[130]

Investor campaigns have some potential to help shift finance away from LAWs, but with several caveats. The presence of a moral villain with intention is "mixed". The private sector is a key part of the emerging regime around LAWS.[131] However, currently there is no definitive list of which companies are involved in the development of LAWs. Three groups are likely to be relevant, which vary in the extent to which they can be portrayed as intentional villains. First are the major defence companies such as Lockheed-Martin, Boeing and BAE Systems. These are also companies that are heavily implicated in the production of nuclear weapons and are thus target companies of the Don't Bank on the Bomb campaign.[132] A second group is technology companies that provide support to these defense companies, either through "translational" work applying AI breakthroughs to a particular military application or through the provision of data storage and computational processing power. For example, Google was involved with Project Maven, a drone AI-imaging program, and had initially bid for the "JEDI" cloud computing contract, both for the Pentagon. Neither was directly tied to LAWS, but are indicative of "applied" work. Third, many of the technologies necessary to create such systems are being produced by technology conglomerates who have no clear incentive to aid the deployment of LAWs. For example, the co-founders of Deepmind (Shane Legg and Demis Hassabis) and a co-founder of OpenAI (Elon Musk) have all signed the pledge on LAWs. Regardless, the AI technologies their companies develop could still enable the weapons they abhor. However, there are also actors contributing to global risk in companies that are directly working on LAWs through military contracts. We thus rate this as a "yes" for the presence of a plausible "villain".

The "hero" for best practice or reinvestment would be AI companies that are not involved with R&D into LAWS. This is not, however, a direct mitigating alternative. Preventative responses are unlikely to be profitable or appealing to institutional investors. Thus, we rate this criterion as "mixed".

Concerns around LAWS have led to calls for a ban on LAWs; the creation of the Campaign to Stop Killer Robots, now supported by 93 non-governmental organisations and 53 countries;[133] United Nations negotiations under the Convention on Certain Conventional Weapons; and a widely-signed pledge by companies, individuals and universities not to participate in the manufacture, use or trade of LAWs.[134] Moreover, the Norwegian Government Pension Fund Global is examining whether companies that contribute to the development of LAWs would be violating fundamental humanitarian norms in doing so.[135] Despite these developments, LAWs have not yet been subject to an investor campaign. However, an investor campaign could draw on the efforts described above, several of the advocates of which have extensive experience with previous arms-control campaigns that included investment tactics.[136]

Investors have significant potential leverage over LAWs. Most defence companies are publicly traded companies in which a few key funds hold significant shares. For example, the five largest holders of shares in Lockheed Martin (State Street Corporation, Capital World Investors, Vanguard Group, BlackRock Inc., and The Bank of America Corporation) cover over a third (41.08%) of total holdings. Together they have substantial leverage over Lockheed's future with LAWs.

At first one might think that investors do not have leverage over technology companies, as they often concentrate voting power in the hands of the founders. However, these same founders have shown themselves to be reactive to public pressure. Investors that engage in contest, protest and request-based tactics are likely to increase leverage. Major technology companies are publicly listed and institutional investors hold significant shares and therefore enjoy some influence. Most tech giants have highly valuable employees who are mindful of ethical challenges and have actively boycotted company activities they disagree with.[137] This was the case when Google decided not to renew its contract with the Pentagon on Project Maven and to withdraw its "JEDI" bid due to employee pushback.[138]

Moreover, Silicon Valley corporates have proven susceptible to shifts in public opinion. In the wake of the Cambridge Analytica scandal, Facebook deployed a suite of responses including public apologies, full-page advertisements in the US and UK to reassure users, a historical audit of data-using apps, and reforms to the accessibility of privacy and security settings.[139] In 2010 Google withdrew from the Chinese market amid public criticism for enabling and legitimising an oppressive regime. A second attempt to launch a censored search engine in China was scuttled in 2018 due to harsh rebukes from free internet advocates, US lawmakers and humanitarian advocates.[140] Together, these factors suggest that AI companies may be susceptible to pressure exerted by institutional investors as well as employees, who are often shareholders themselves.

Accordingly, we rate LAWs as "mixed" for the presence of a villain with intention, "mixed" for the presence of a clear alternative, "no" for the presence of a current campaign, and "high" for investor leverage. If particular companies are clearly involved in the development of LAWS, an investor campaign might well be appropriate and successful.

7.3 Advanced Artificial Intelligence (AI) systems

As AI systems become more powerful, and as our societies become more reliant on them, the risks we face will also increase. The field of AI is advancing rapidly,[141] but there is uncertainty about the speed of future progress.[142] Yet even existing systems have raised multiple concerns around issues such as labour automation, algorithmic bias and privacy intrusions. Longer-term fears include the reinforcement of authoritarian regimes,[143] new physical, political and cybersecurity concerns,[144] an arms race,[145] a destabilisation of the global geopolitical order,[146] or even a failure of nuclear deterrence AI systems.[147] Even well-designed systems could trigger these accidents due to their sheer complexity and tightly coupled nature.[148]

The nascency of many of the technologies makes rating advanced AI systems difficult. They are likely to be "mixed" on the presence of a villain and hero. AI systems are often dual use, and the same companies researching dangerous applications may also be developing beneficial ones; in many cases the hero and villain may be one and the same. For instance, Google Deepmind and OpenAI are at the forefront of pioneering intelligent AI systems, but are also at the forefront of AI safety research. They have both released sets of ethical principles regarding AI.[149] In 2019, OpenAI even resorted to withholding the source code of a language-processing algorithm it created due to concerns over potential malicious misuses.[150] Yet, one might question why OpenAI was even developing such a reckless technology in the first place. As this distinction between responsible and irresponsible developers becomes clearer in the future, the ability to act against the irresponsible will become more urgent and easier to do.

To date, no campaign has suggested investment redirection as a way to avoid the creation of dangerous AI systems. However, an investor campaign could draw on the existing "AI safety" community.[151] We have rated advanced AI systems as "mixed" for the presence of a compelling villain, "mixed" for the presence of a clear alternative hero for reinvestment, "no" for a campaign, and "medium" for investor leverage. If in the future companies were to develop advanced AI in an irresponsible manner, an investor campaign might be appropriate and successful.

7.4 Asteroid strikes

Investor campaigns are least applicable to asteroid strikes. There is no clear villain with intention. No corporation, state or other organisation is responsible for causing asteroids to collide with the Earth. Because of this, investors have no targets with tangible assets to influence. No campaign has yet been enacted for the public divestment from asteroid strikes or other natural risks, for obvious reasons. However, there are clear mitigating technologies that funds could finance. This includes programs for the detection of near-Earth objects (NEOs) and preventative responses such as asteroid deflection projects.

Most of these mitigating approaches are unlikely to be profitable for private companies and are largely under the purview of states. However, there may be some opportunities for profitable commercial investments.

For example, while the monitoring of NEOs is largely done by the US National Space Agency's (NASA) Planetary Defense Coordination Office, part of their NEO tracking efforts are performed by the NEOWISE satellite. The satellite was built by Ball Aerospace and Technologies, a subsidiary of the Ball Corporation, a publicly traded company.[152]

There is one tension in using finance to mitigate asteroid risk. Deflecting an incoming NEO may require the use of a nuclear warhead. Risk-risk trade-off analyses of the tension between disarmament and maintaining a nuclear arsenal for asteroid deflection have yet to come to a clear conclusion.[153] This is further complicated by such use of nuclear weapons being prohibited under internal law. However, this could be circumvented through the adoption of a resolution under the UN General Assembly.[154] For now, it appears that at least the act of preventing the modernisation and growth of nuclear arsenals is compatible with both reducing the risk of NEOs and nuclear weapons. Such trade-offs for global risk investor campaigns require closer attention and research.

Most of the key points here are applicable to other natural risks such as super volcanoes and comet strikes. Humans are not responsible for these hazards; there is no actor to deter or divest from. However, finance can be redirected towards endeavours that improve our resilience to these threats. While many of these will fall under the mandate of states, there may be some lucrative options for mitigating technologies, especially if any of these risks grow in public profile.

7.5 Dealing with dual use

For many technological applications it is too early to contest, protest, request, divest, reinvest or acquest. In other cases, a clear stand can be taken. For instance, investors might oppose technologies being sold to governments for the purpose of sustaining an authoritarian surveillance state. This is already a live issue for relationships with China. To date Yahoo has provided the Chinese government with information on a journalist that allowed him to be jailed, Cisco helped install 500,000 cameras in Chongqing and Thermo Fisher Scientific supplied the Chinese government with DNA sequencers that have been used to target ethnic minorities in Xinjiang.[155] Yet these clear-cut cases are not the norm. Many technologies that might be transformative are still in their nascency. For now, it is difficult to tell whether they will be developed in socially beneficial or harmful ways. Once developed they are likely to remain "dual use": the same technology may be capable of being used for benevolent or malicious purposes. An alternative approach is needed for these dual-use systems.

We suggest that a seventh tactic of "*engagement*" is most appropriate for developing dual-use systems. In this case a new investor campaign would identify demand and supply watchlists. The supply watchlist would contain companies that could be at high risk of developing harmful dual-use technologies. This would be accompanied by a watchlist of actors that are likely to put dual-use technologies to malicious use. This could include actors offering contracts, future purchasing agreements or funding for harmful technologies. Both watchlists would be underpinned by a review of relevant information on the technology, including potential uses and misuses, the components underlying it and the materials and knowledge required to manufacture it. The information catalogue would help to characterise dual-use concerns and to determine how the production of a dual-use technology can be separated out from the wider operations of a company.

The information catalogue and demand and supply watchlists can then provide the basis for engagement with relevant companies. The aim of this dialogue would be to generate a set of mutually agreed guidelines on the development, deployment and sale of dual-use technologies. Such guidelines could include:

- Publicly disclosing all contracts and funding in areas of concern;
- Agreeing to a black-list of potential collaborators due to dual-use concerns;
- Publicly releasing ethical guidelines on the research, development and deployment of dual-use technologies; and
- Hosting special sessions during board meetings to discuss dual-use concerns.

These guidelines could provide a filter for sorting responsible from irresponsible corporate actors. Businesses that agree to the guidelines and uphold them can be marked as safe investments. Those that do not adhere to the guidelines can be subject to escalation tactics. Such an escalation could run the gauntlet from contest through to request and acquest tactics. An overview of this engagement approach is outlined in Figure 2. Ongoing campaigns are already approximately following this process. PAX has enshrined a new campaign against LAWs in a series of reports.[156] These provide information on the technology and its potential misuses as well as a list of companies involved in their development along with a rating of their current performance according to the standards of "best practice", "medium concern", and "high concern". In a series of recommendations, it has also set forth guidelines and suggested processes including a public commitment to avoid working on LAWs, internal policies and informing employees of the company's relationship to LAWs development. All of this has preceded any escalation into a distinct investor campaign. PAX is unknowingly already following the framework for engagement.

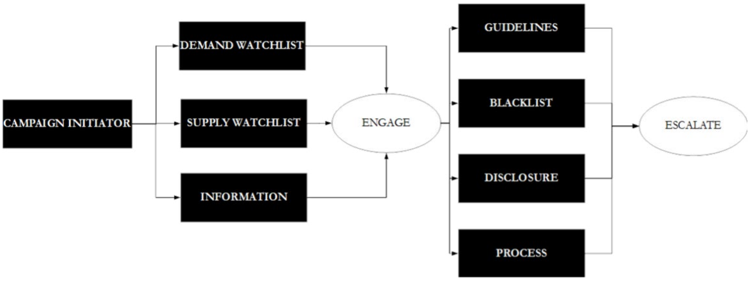


Fig. 2: A process for dealing with dual-use technologies.

Conclusions: High Money, Higher Stakes

Investor campaigns seek to change corporate activities or redirect finance and turn industrial titans into subjects of public policy. These campaigns are gaining traction as a strategy for addressing climate change and the threat of nuclear weapons but have largely been overlooked for other global risks. This should not be so. The ethical arguments used to justify campaigns against Apartheid South Africa, tobacco and fossil fuels can be applied to the ambit of global hazards. They are a significant moral wrong that investors should be compelled to prevent. The applicability of these motivations to other risks varies according to actor and issue. Legal and financial obligations differ across countries and markets. Some institutions may have an inbuilt goal of perpetuity and long-term survival, or may be universal owners with a bias towards maintaining a healthy global economy. Others may not. Divestment has been either profitable or at the very least posed little cost in some cases, such as fossil fuels. In others, such as tobacco, it may have led to foregone profits.

Our analysis yields two key conclusions about why institutional investors should use such campaigns. First, institutional investors have strong ethical and other grounds to avoid contributing to global risks. At a minimum they should commit to a Financial Hippocratic Oath (the traditional oath for doctors, requiring them to "First do no harm"): a core principle to not contribute to global risk through their investments. Second, institutional investors who are universal owners, perpetual institutions or have significant exposure to Global Catastrophic Risks should commit to a Financial Oath of Maimonides (the traditional oath for physicians and pharmacists that is more expansive in its requirements, opening with "The eternal providence has appointed me to watch over the life and health of thy creatures"): a pledge to use their investments to protect the world from global risks and create widespread social benefits.

An investor campaign is a tried-and-tested method for exposing socially harmful actions to public scrutiny and oversight. Such campaigns are helping to remedy global ills ranging from nuclear weapons to climate change, yet their potential has not been fully tapped. Global risks are the next realm for institutional investors, activist shareholders, and investment campaigners to tackle. The reasons for institutional investors to consider global risks are compelling. Together, they provide strong grounds for campaigners to begin to tackle other emerging sources of catastrophe aside from climate change and nuclear war.

In terms of how institutional investors can do this; different tactics become appropriate depending on an institutional investor's holdings and the campaign target's ability and willingness to change. Investor campaigns are only likely to be fruitful in cases in which there is a compelling and intentional moral "villain", a clear alternative available, a galvanising campaign and sufficient investor leverage. Few global risks fully fit this profile. Most extreme technological risks, such as advanced AI systems and catastrophic biotechnological risks, are emerging and dual use.

For these emerging dual-use technologies with global risk potential, a more nuanced tactic of engagement is needed. This would allow cooperation rather than confrontation to mark the relationship between institutional investors and the actors developing these technologies. The tactic of engagement hinges on informed discussions between investors and corporations leading to guidelines that can help separate responsible from irresponsible actors. It also provides a measure against which corporate actions can be monitored and assessed. Such a detailed and cautious tactic is needed to ensure that companies steer technological development in socially beneficial, rather than risky, directions.

Our analysis suggests that investment campaigns have confronted risks that currently fit their tactics. The choice may have been unconscious, but is nonetheless astute. Both climate change and nuclear war have clear villains with an intention to cause harm (or at least knowledge of these harms) and substantial investor leverage. Climate change has both higher investor leverage (due to a higher reliance on private rather than state companies than is the case for nuclear weapons) and, unlike nuclear weapons, clear alternatives. This is potentially why the Fossil Free movement has experienced greater coverage and traction than the Don't Bank on the Bomb campaign.[157]

Many of the most precipitous risks looming on humankind's horizon are being crafted and funded by a few actors. Investor campaigns provide one way for concerned investors and activists to together steer the world towards safety. Investors operating under a Financial Hippocratic Oath and Financial Oath of Maimonides have an array of tactics to draw on, and many are applicable to the emerging threats of the future. These tactics can help ensure that we are no longer collectively financing our final hour.

Acknowledgements

This project was made possible through the support of a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of Templeton World Charity Foundation. We would like to thank Frank Rövekamp as well as Maaike Beenes and Daan Kayser for their helpful comments on an earlier draft of this chapter. We also extend our thanks to our colleagues at CSER for their astute advice on earlier versions, and to several anonymous investors for their comments. Last but not least we would like to thank Leon Ritter for his assistance in helping to revise the chapter for submission.

Notes and References

1 Raji, M. Y. 'Timeline: Fossil fuels divestment', *The Crimson* (2 October 2014). http://www.thecrimson.com/article/2014/10/2/timeline-fossil-fuels-divestment/

2 *Global Fossil Fuel Divestment Commitments Database* (2022). https://divestmentdatabase.org/

3 ICAN. *Don't Bank on the Bomb* (2018). https://www.dontbankonthebomb.com/

4 World Economic Forum (WEF). *The Global Risks Report 2020* (2020). https://www.weforum.org/publications/the-global-risks-report-2020

5 Garschagen, M., S. L. R. Wood, J. Garard, M. Ivanova and A. Luers. 'Too big to ignore: Global risk perception gaps between scientists and business leaders', *Earth's Future, 8* (2020):1–5. https://doi.org/10.1029/2020EF001498; Future Earth. *Future Earth Risks Perceptions Report 2020* (2020):1–13. https://futureearth.org/initiatives/other-initiatives/grp/the-report/

6 Global Challenges Foundation. *Global Catastrophic Risks Report 2018*. Global Challenges Foundation (2018). https://www.humanitarianfutures.org/global-catastrophic-risk-report-2018/; Atkinson, A. *Impact Earth: Asteroids, Comets, and Meteors — The Growing Threat*. Virgin (2019)

7 Robock A., L. Oman and G. L. Stenchikov. 'Nuclear winter revisited with a modern climate model and current nuclear arsenals: Still catastrophic consequences', *Journal of Geophysical Research: Atmospheres, 112*(D13) (2007): 1–14. https://doi.org/10.1029/2006jd008235; Borrie, J., T. Caughley and W. Wan (eds.). *Understanding Nuclear Weapon Risks*. UNIDIR (2017).

8 Weitzman, M. 'On modeling and interpreting the economics of catastrophic climate change', *The Review of Economics and Statistics, 91*(1) (2009): 1–19. https://doi.org/10.1162/rest.91.1.1; Xu, Y. and V. Ramanathan. 'Well below 2 °C: Mitigation strategies for avoiding dangerous to catastrophic climate changes', *Proceedings of the National Academy of Sciences, 114*(39) (2017): 10315–23. https://doi.org/10.1073/pnas.1618481114

9 Barrett, A. M., S. D. Baum and K. Hostetler. 'Analyzing and reducing the risks of inadvertent nuclear war between the United States and Russia', *Sci. Glob. Secur., 21* (2013): 106–33. https://doi.org/10.1080/08929882.2013.798984

10 Wagner, G. and M. L. Weitzman. *Climate Shock*. Princeton University Press (2016).

11 Taleb, N. N. *The Black Swan: The Impact of the Highly Improbable* (Vol. 2). Random House (2007).

12 Betz, G. 'Accounting for possibilities in decision making', in S. O. Hansson and H. G. Hadorn (eds.), *The Argumentative Turn in Policy Analysis*. Springer (2016). pp. 135–69.

13 Ord, T. *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury Press (2020).

14 Kaczmarek P, Beard S. Human Extinction and Our Obligations to the Past. *Utilitas*. 2020;32(2):199–208. https://doi.org/10.1017/s0953820819000451

15 Beard S, Kaczmarek P. On the Wrongness of Human Extinction. 2020 Argumenta https://doi.org/10.14275/2465-2334/20199.bea

16 Centeno, M. A., M. Nag, T. S. Patterson, A. Shaver and A. J. Windawi. 'The emergence of global systemic risk', *Annu. Rev. Sociol., 41* (2015): 65–85. https://dx.doi.org/10.1146/annurev-soc-073014-112317

17 Homer-Dixon, T., B. Walker, R. Biggs, A. S. Crépin, C. Folke, E. F. Lambin, G. D. Peterson, J. Rockström, M. Scheffer, W. Stef-fen and M. Troell. 'Synchronous failure: The emerging causal architecture of global crisis', *Ecology and Society, 20*(3) (2015):1–16. https://doi.org/10.5751/es-07681-200306

18 Beck, U. 'From industrial society to the risk society: Questions of survival, social structure and ecological enlightenment', *Theory, Culture & Society, 9*(1) (1992): 97–123. https://doi.org/10.1177/026327692009001006

19 Griffin, P. *The Carbon Majors Database: CDP Carbon Majors Report 2017*. CDP and the Climate Accountability Institute (2017).

20 ICAN (2018).

21 Slijper, F., A. Beck and D. Kayser. 'State of AI: Artificial Intelligence, the military and increasingly autonomous weapons', *PAX* (2019a).

22 Cremer, C. Z. 'Deep limitations? Examining expert disagreement over deep learning', *Progress in Artificial Intelligence* (2021).

23 Fitzgerald, M., A. Boddy and S. D. Baum. '2020 survey of Artificial General Intelligence projects for ethics, risk, and policy', *Global Catastrophic Risk Institute Technical Report, 20*(1) (2020).

24 Fitzgerald et al. (2020); Baum, S. et al. 'A survey of Artificial General Intelligence projects for ethics, risk, and policy', *Global Catastrophic Risk Institute Working Paper, 17*(1) (2017).

25 TERRA. *The Existential Risk Research Assessment* (*TERRA*) (2021). terra.cser.ac.uk

26 Shackelford, G., L. Kemp, C. Rhodes, L. Sundaram, S. ÓhÉigeartaigh and S. Beard et al. 'Accumulating evidence using crowdsourcing and machine learning: A living bibliography about existential risk and Global Catastrophic Risk', *Futures, 116* (2020): 1–10. https://doi.org/10.1016/j.futures.2019.102508

27 Krysiak, F. 'Risk management as a tool for sustainability', *Journal of Business Ethics, 85* (2009): 483–92. https://doi.org/10.1007/s10551-009-0217-7

28 Galaz, V., B. Crona, A. Dauriach, B. Scholtens and W. Steffen. 'Finance and the Earth system — Exploring the links between financial actors and non-linear changes in the climate system', *Global Environmental Change, 53* (2018): 296–302. https://doi.org/10.1016/j.gloenvcha.2018.09.008

29 Galaz et al. (2018); Keys, P. W., V. Galaz, M. Dyer, N. Matthews, C. Folke, M. Nyström and S. E. Cornell. 'Anthropocene risk', *Nature Sustainability, 2* (2019): 667–73.

30 Green, F. 'Anti-fossil fuel norms', *Climatic Change, 150*(103) (2018).

31 Ayling, J. and N. Gunningham. 'Non-state governance and climate policy: The fossil fuel divestment movement', *Climate Policy, 17*(2) (2017): 131–49. https://doi.org/10.1080/14693062.2015.1094729; Piggot, G. 'The influence of social movements on policies that constrain fossil fuel supply', *Climate Policy* (2018). https://doi.org/10.1080/14693062.2017.1394255

32 Ayling, J. 'A contest for legitimacy: The divestment movement and the fossil fuel industry', *Law and Policy, 39*(4) (2017): 349–71. https://doi.org/10.1111/lapo.12087

33 Triodos Bank and Ethical Consumer. *Ethical Consumer: Markets Report 2017* (2017).

34 Hunt, C., O. Weber and T. Dordi. 'A comparative analysis of the anti-Apartheid and fossil fuel divestment campaigns', *Journal of Sustainable Finance and Investment, 7*(1) (2018): 64–81. https://doi.org/10.4324/9780203701164-8

35 Ansar, A., B. Caldecott and J. Tilbury. 'Stranded assets and the fossil fuel divestment campaign: What does divestment mean for the valuation of fossil fuel assets?', *Technical Report*. Smith School of Enterprise and the Environment (2013).

36 Fihn, B. 'The logic of banning nuclear weapons', *Survival, 59*(1) (2017): 43–50. https://doi.org/10.1080/00396338.2017.1282671

37 Bergman, N. 'Impacts of the fossil fuel divestment movement: Effects on finance, policy and public discourse', *Sustainability, 10*(2519) (2018): 1–18. https://doi.org/10.3390/su10072529

38 Ansar, Caldecott and Tilbury (2013).

39 Tarrow, S. 'Modular collective action and the rise of the social movement: Why the french revolution was not enough', *Politics and Society, 21* (1993): 69–90. https://doi.org/10.1177/0032329293021001004; Tilly, C. 'Contentious repertoires in Great Britain, 1758–1834', *Social Science History, 17*(1993): 253–80. https://doi.org/10.2307/1171282

40 Soule, S. A. 'The student divestment movement in the United States and tactical diffusion: The shantytown protest', *Social Forces*, 75(3) (1997): 855–82. https://doi.org/10.2307/2580522

41 Hiltzik, M. 'When is it worth it to divest?', *Los Angeles Times* (January 2016).

42 Slijper et al. (2019).

43 Beenes, M. and S. Snyder. *Don't Bank on the Bomb 2018: A Global Report on the Financing of Nuclear Weapons Producers* (2018).

44 Ayling, J. and N. Gunningham. 'Non-state governance and climate policy: The fossil fuel divestment movement', *Climate Policy, 17*(2) (2017): 131–49. https://doi.org/10.1080/14693062.2015.1094729

45 McKibben, B. 'Global warming's terrifying new math', *Rolling Stone* (July 2012). https://www.rollingstone.com/politics/politics-news/global-warmings-terrifying-new-math-188550/

46 Lenferna, A. 'Divest — invest: A moral case for fossil fuel divestment', in R. Kanbur and H. Shue (eds.). *Climate Justice: Integrating Economics and Philosophy*. Oxford University Press (2018). pp. 139–56.

47 Hofferberth, M., T. Brühl, E. Burkart, M. Fey and A. Peltner. 'Multinational enterprises as "social actors": Constructivist explanations for corporate social responsibility', *Global Society, 25*(2) (2010): 205–26. https://doi.org/10.1080/13600826.2011.553533

48 Davis, J. H., F. D. Schoorman and L. Donaldson. 'Toward a stewardship theory of management', *Academy of Management Review, 22* (1997): 20–47. https://doi.org/10.2307/259223

49 Caldwell, C., L. A. Hayes, P. Bernal et al. 'Ethical stewardship – Implications for leadership and trust', *Journal of Business. Ethics, 78* (2008): 153–64. https://doi.org/10.1007/s10551-006-9320-1

50 Mueller, Tom. *Crisis of Conscience: Whistleblowing in an Age of Fraud*. Riverhead (2019); Schoen, E. J. 'The 2007–2009 financial crisis: An erosion of ethics: A case study', *Journal of Business Ethics, 146* (2017): 805–30. https://doi.org/10.1007/s10551-016-3052-7

51 Bank of England Prudential Regulation Authority. 'Enhancing banks' and insurers' approaches to managing the financial risks from climate change', *Consultation Paper, 23*(18) (2018).

52 ClientEarth. 'EasyJet among companies reported to regulator by ClientEarth',

ClientEarth (2018). https://www.clientearth.org/easyjet-among-companies-reported-to-regulator-by-clientearth/

53 Del Guercio, D. 'The distorting effect of the prudent-man laws on institutional equity investments', *Journal of Financial Economics, 40*(1) (1996): 31–62. https://doi.org/10.1016/0304-405x(95)00841-2

54 Merkt, H. 'Rechtliche Grundlagen der Business Judgment Rule im internationalen Vergleich zwischen Divergenz und Konvergenz', *Zeitschrift für Unternehmens- und Gesellschaftsrecht, 46*(2) (2017): 129–48.

55 Schanzenbach, M. M. and R. H. Sitkoff. 'Fiduciary duty, social conscience, and ESG investing by a trustee', *55 Annual Heckerling Institute on Estate Planning*, ed. Tina Portando (2021).

56 Freshfields Bruckhaus Deringer. *A Legal Framework for Impact: Sustainability Impact in Investor Decision-Making* (2021).

57 Stout, L. *The Shareholder Value Myth: How Putting Shareholders First Harms Investors, Corporations, and the Public*. Berrett-Koehler Publishers (2012).

58 *Charter of Newnham College, Cambridge* (1917). https://www.newn.cam.ac.uk/wp-content/uploads/2015/08/Charter-Statutes.pdf

59 The Archbishops' Council. *Annual Report and Financial Statements for the Year Ended 31. December 2012* (May 30 2013). https://www.churchofengland.org/sites/default/files/2018-01/gs%201913%20-%20ac%20report_July13.pdf

60 Charity Commission. *Permanent Endowment: Rules for Charities* (2014). https://www.gov.uk/guidance/permanent-endowment-rules-for-charities#history

61 Norges Bank Investment Bank. *Mission and Values* (22 May 2017). https://www.nbim.no/en/organisation/career/our-culture/mission-and-values/

62 Trinks, A. et al. 'Fossil fuel divestment and portfolio performance', *Ecological Economics, 146* (2018): 740–48.

63 Grantham, J. 'The race of our lives revisited', *GOM White Paper* (2018).

64 Bello, Z. 'Socially responsible investing and portfolio diversification', *Journal of Financial Research, 28*(1) (2005): 41–57; Humphrey, J. E. and D. T. Tan. 'Does it really hurt to be responsible?', *Journal of Business Ethics, 122*(3) (2014): 376–86; Lobe, S. and C. Walkshäusl. 'Vice versus virtue investing around the world', *Review of Managerial Science, 10* (2016): 303–44.

65 Halcoussis, D. and A. D. Lowenberg. 'The effects of the fossil fuel divestment campaign on stock returns', *North American Journal of Economics and Finance* (July 2018): 669–74.

66 Dimson E., O. Karakaş and X. Li. 'Active ownership', *The Review of Financial Studies, 28*(12) (2015): 3225–68.

67 Yu, L. 'Performance of socially responsible mutual funds', *Global Journal of Business Research, 6* (2014): 9–17.

68 Halcoussis and Lowenberg (2018).

69 Dimson, E., P. Marsh and M. Staunton. 'Industries: Their rise and fall', *Credit Suisse Global Investment Returns Yearbook 2015* (2015).

70 Lewis, M. *The Big Short: Inside the Doomsday Machine*. W.W Norton and Company (2010).

71 McKibben (2012).

72 Carbon Tracker. *Unburnable Carbon: Are the World's Financial Markets Carrying a Carbon Bubble*? Carbon Tracker (2011).

73 McGlade, C. and P. Ekins. 'The geographical distribution of fossil fuels unused when limiting global warming to 2 °C', *Nature, 517*(7533) (2015): 187–90.

74 Ayling and Gunningham (2017).

75 Mercure, J. F., H. Pollitt, J. E. Viñuales, N. R. Edwards, P. B. Holden and U. Chewpreecha et al. 'Macroeconomic impact of stranded fossil fuel assets', *Nature Climate Change, 8* (2018): 588–93.

76 Cambridge Institute for Sustainability Leadership. *Unhedgeable Risk: How Climate Change Sentiment Impacts Investment* (2015).

77 Lewis, S. C., S. E. Perkins-Kirkpatrick, G. Althor, A. D. King and L. Kemp. 'Assessing contributions of major emitters' Paris-era decisions to future temperature extremes', *Geophysical Research Letters, 46*(2019): 1–8; Schiermeier, Q. 'Droughts, heatwaves and floods: How to tell when climate change is to blame', *Nature, 560* (2018): 20–22; Nature. 'Pinning extreme weather on climate change is now routine and reliable', *Nature, 560*(5) (2018); Gallant, A. and S. Lewis. 'Stochastic and anthropogenic influences on repeated record-breaking temperature extremes in Australian spring of 2013 and 2014', *Geophysical Research Letters, 43* (2016): 2182–91; Dittus, A. et al. 'A multiregion assessment of observed changes in the areal extent of temperature and precipitation extremes', *Journal of Climate, 28* (2015): 9206–20.

78 Ekwurzel, B., J. Boneham, M. W. Dalton, R. Heede, R. J. Mera, M. R. Allen and P. C. Frumhoff. 'The rise in global atmospheric CO2, surface temperature, and sea level from emissions traced to major carbon producers', *Climatic Change, 144* (2017): 579–90.

79 Mooney, A. and E. Crooks. 'New York sues big oil companies over climate change', *Financial Times* (10 January 2018).

80 Watts, J. 'Shell threatened with legal action over climate change contributions‘, *The Guardian* (4 April 2018).

81 DiChristopher, T. 'Judges tosses out NYC climate change lawsuit against 5 major oil companies', *CNBC* (19 July 2018).

82 Roose, K. 'Why napalm is a cautionary tale for tech giants pursuing military contracts', *New York Times* (4 March 2019). https://www.nytimes.com/2019/03/04/technology/technology-military-contracts.html

83 Darling, K., J. Arm and R. Gatlin. 'How to effectively reward employees', *Industrial Management* (July/August 1997): 1–4; Turner, J. H. 'Pay for performance: Contrary evidence and a predictive model', *Academy of Marketing Studies Journal, 10* (2006): 23–38.

84 Montgomery, D. B. and C. A. Ramus. 'Calibrating MBA job preferences for the 21st century', *Academy of Management Learning & Education, 10*(1) (2011: 9–26.

85 Woodruffe, C. 'A potent secret for winning a crucial edge over your rivals?', *Industrial & Commercial Training, 38*(1) (2006): 18–22.

86 Urwin, R. 'Pension funds as universal owners: Opportunity beckons and leadership calls‘, *Rotman International Journal of Pension Management, 4*(1) (2011): 26–34.

87 Hawley, J. P. and A. Williams. 'The universal owner's role in sustainable economic development', *Corporate Environmental Strategy, 9*(39) (1997): 284–91; Hawley, J. and A. Williams. 'Universal owners: Challenges and opportunities', *Corporate Governance: An International Review, 15*(3) (2007): 415–20.

88 Gjessing, O. P. K. and H. Syse. 'Norwegian petroleum wealth and universal ownership', *Corporate Governance: An International Review, 15*(3) (2007): 427–37; Lydenberg, S. 'Universal investors and socially responsible investors: A tale of emerging affinities', *Corporate Governance: An International Review, 15*(3) (2007): 212–17.

89 Council on Ethics for the Norwegian Government Pension Fund Global. *Annual Report 2017* (2018).

90 Rockefeller Brothers Fund. *Fossil Fuel Divestment* (2019). https://www.rbf.org/mission-aligned-investing/divestment

91 Winston, Morton. 'NGO strategies for promoting corporate social responsibility', *Ethics and International Affairs, 16*(2) (2002): 71–87.

92 Winston (2002).

93 O'Rourke, A. 'A new politics of engagement: Shareholder activism for corporate social responsibility', *Business Strategy and the Environment, 12*(4) (2003): 227–39. Gillan, S. L. and L. T. Starks. 'The evolution of shareholder activism in the United States', *Journal of Applied Corporate Finance, 19*(1) (2007): 55–73.

94 This often is not directly effective in shifting company behaviour. Many companies will not listen to shareholders who do not hold a substantial amount of stock. However, the more frequent goal is to *publicly* humiliate and critique corporations in order to change their behaviour or undermine their social licence. This is why these protest actions are often taken by activists and leaked to the media.

95 Investor Agenda. *Policy Advocacy* (2019). https://theinvestoragenda.org/areas-of-impact/policy-advocacy/

96 "Reinvest" does not have impact if it involves the purchase of shares that are already listed on the stock market; reinvestment should involve allocating capital to companies, cooperatives or projects contributing to risk-attenuating activities, not using capital to purchase shares from fellow shareholders. For example, an investor can "steer financing away" from fossil fuels by reinvesting divested funds into asset classes other than public equity. Divestments within public equity arguably do not technically "steer financing" as these assets are already listed on the stock market.

97 One example of this is the Australian superannuation fund Future Super. *Future Super: Empowering Everyday Aussies to Invest in a Renewable Future* (2018). https://www.myfuturesuper.com.au/. The fund offers both a "Balanced Growth" option which is divested from fossil fuels, as well as a "Renewables Plus" option that is fossil free and targets a fifth of the portfolio towards renewable energy projects.

98 Winston (2002).

99 This included interviews with an official working on the Don't Bank on the Bomb Campaign, another on the Campaign to Stop Killer Robots, two personnel members with the Humane League, two former tech employees, two workers with PAX, an employee with Baillie Gifford, and an official with the United Nations Institute for Disarmament Research (UNIDR).

100 Oreskes, N. and E. Conway. *Merchants of Doubt: How a Handful of Scientists Obscured the Truth on Issues From Tobacco Smoke to Global Warming*. Bloomsbury (2010).

101 InfluenceMap. 'Big oil's real agenda on climate change', *InfluenceMap* (2019). https://influencemap.org/report/How-Big-Oil-Continues-to-Oppose-the-Paris-Agreement-38212275958aa21196dae3b76220bddc

102 Quigley, E. *Universal Ownership in the Anthropocene* (2020). https://ssrn.com/abstract=3457205 or http://dx.doi.org/10.2139/ssrn.3457205

103 Belfield, H. 'The machine learning community as a political actor', *AIES 2020 Conference* (2019).

104 Benford, R. D. and S. A. Hunt. 'Dramaturgy and social movements: The social construction and communication of power', *Sociological Inquiry, 62*(1) (1992): 36–55.

105 Jones, M. D. and M. K. McBeth. 'A narrative policy framework: clear enough to be wrong?', *Journal of Policy Studies, 38*(2) (2010): 329–53.

106 Marshall, G. *Don't Even Think About It: Why Our Brains are Wired to Ignore Climate Change*. Bloomsbury (2018); Lorenzoni, I., S. Nicholson-Cole and L. Whitmarsh. 'Barriers perceived to engaging with climate change among the UK public and their policy implications', *Global Environmental Change, 4* (2007): 445–59.

107 Jones, M. D. 'Cultural characters and climate change: How heroes shape our perception of climate science', *Social Sciences Quarterly, 95*(1) (2014): 1–39.

108 Missile defence systems might be considered as a mitigating alternative, but suffer from multiple drawbacks. First, the major investors are governments, not corporations. Second, many of the corporations providing enabling technologies are also nuclear or arms producers. Third, successful missile defence systems could have perverse consequences. Being protected from attacks could incentivise first-use policies.

109 Beenes and Snyder (2018).

110 Dimson, Karakaş and Li (2015).

111 Winston (2002).

112 Oliver, P. E. and G. Marwell. 'The paradox of group size in collective action: a theory of the critical mass', *American Sociological Review* (1988): 1–8.

113 For further discussion of the distinction between primary and secondary market ESG activities and effectiveness, see Quigley (2020).

114 One could also consider to what extent the tactics could be applied to other potential global risks, such as Solar Radiation Management, catastrophic ecosystem shifts or natural pandemics.

115 The above subjective estimates are based on the authors' judgements, which we justify below. Some of the criteria, such as intention and the existence of a moral villain, are by nature qualitative and subjective. Others, such as the amount of stocks held by institutional investors, are quantitative questions that could be uncovered through further research.

116 Rees, M. J. *Our Final Century*? William Heinemann (2003).

117 Koblentz, G. D. 'The de novo synthesis of horsepox virus: Implications for biosecurity and recommendations for preventing the reemergence of smallpox', *Health Security, 15*(6) (2017): 620–8.

118 Schoch-Spana, M., A. Cicero, A. Adalja, G. Gronvall, T. Kirk Sell, D. Meyer, J. B. Nuzzo, S. Ravi, M. P. Shearer, E. Toner, C. Watson, M. Watson ... T. Inglesby. 'Global Catastrophic Biological Risks: Toward a working definition', *Health Security, 15*(4) (2017): 323–28; Millett, P. and A. Snyder-Beattie. 'Existential risk and cost-effective biosecurity', *Health Security, 15*(4) (2017): 373–83.

119 That is, comparable in destructiveness to nuclear weapons.

120 Carus, W. S. 'A century of biological-weapons programs (1915–2015): Reviewing the evidence', *The Nonproliferation Review, 24*(1–2) (2017): 129–53.

121 Danzig, R., M. Sageman, T. Leighton, L. Hough, H. Yuki, R. Kotani and Z. M. Hosford. *Aum Shinrikyo: Insights onto How Terrorists Develop Biological and Chemical Weapons*. Centre for a New American Security (2012).

122 Herfst, S. et al. 'Airborne transmission of influenza A/H5N1 virus between ferrets', *Science, 336*(6088) (2012): 1534–41; Imai, M. et al. 'Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferret', *Nature, 486*(7403) (2012): 420–28. The transmissibility of the virus was not tested directly; rather, it was manipulated so as to be transmissible between ferrets, which are the standard model for human transmissibility.

123 Lipsitch, M. and T. V. Inglesby. 'Moratorium on research intended to create novel potential pandemic pathogens', *mBio, 5*(6) (2015): 1–6; Farquhar, S. and O. C.-B. A. Snyder-Beattie. 'Pricing externalities to balance public risks and benefits of research', *Health Security, 15*(4) (2017): 401–08.

124 An investor campaign might also become appropriate if there were evidence that the protocol was too weak, or that one of the companies of the IGSC was flouting the protocols or in some other manner contributing to Global Catastrophic Biological Risks. To be clear, we have no such evidence currently and we believe the IGSC is a model of an industry taking responsibility and acting collaboratively.

125 Marchant, G. E. et al. 'International governance of autonomous military robots', *Science and Technology Review, 12*(1) (2011): 272–315.

126 Arkin, R. 'Lethal autonomous systems and the plight of the non-combatant', *AISB Quarterly, 137* (2013): 1–9.

127 Boulanin, V. and M. Verbruggen. *Mapping the Development of Autonomy in Weapon Systems*. Stockholm International Peace Research Institute (2017).

128 Future of Life Institute. *Autonomous Weapons: An Open Letter From AI and Robotics Researcher*s (2015). https://futureoflife.org/open-letter/open-letter-autonomous-weapons-ai-robotics/

129 Russell, S., A. Aguirre, A. Conn and M. Tegmark. 'Why you should fear "slaughterbots" — A response', *IEEE Spectrum* (2018). https://spectrum.ieee.org/automaton/robotics/artificial-intelligence/why-you-should-fear-slaughterbots-a-response

130 Scharre, P. 'Autonomous weapons and operational risk', *Ethical Autonomy Project. 20YY Future of Warfare Initiative* (2016a). Center for a New American Security. https://s3.amazonaws.com/files.cnas.org/documents/CNAS_Autonomous-weapons-operational-risk.pdf; Scharre, P. *Flash War — Autonomous Weapons and Strategic Stability*. Presented at the Understanding Different Types of Risk, Geneva (2016b). http://www.unidir.ch/files/conferences/pdfs/-en-1-1113.pdf. Some also object to LAWS on deontological grounds, arguing that a machine should never "decide" to kill a human.

131 Frederick, K. 'The civilian private sector: Part of a new arms control regime?', in R. Passi. (ed.). *Raisina Files Volume 4: Debating Future Frameworks in a Disrupted World*. Observer Research Foundation (2019).

132 Beenes and Snyder (2018).

133 Campaign to Stop Killer Robots. *Campaign to Stop Killer Robots* (2018). https://www.stopkillerrobots.org/

134 Future of Life Institute. *Lethal Autonomous Weapons Pledge* (2018). https://futureoflife.org/lethal-autonomous-weapons-pledge/?cn-reloaded=1

135 Goose, S. 'The growing international movement against killer robots', *Harvard International Review, 37*(4) (2017): 28–33.

136 For example, Mary Wareham, coordinator of the Campaign to Stop Killer Robots, worked for the Vietnam Veterans of America Foundation in support of its Nobel prize-winning International Campaign to Ban Landmines, and for Oxfam New Zealand, leading its efforts to secure an arms trade treaty and the 2008 *Convention on Cluster Munitions — Human Rights Watch*. Mary Wareham profile (2019). https://www.hrw.org/about/people/mary-wareham. Accessed 18 February 2019

137 Belfield (2019).

138 Conger, Kate. 'Google plans not to renew its contract for project maven, a controversial drone AI imaging program', *Gizmodo* (2018). https://gizmodo.com/google-plans-not-to-renew-its-contract-for-project-mave-1826488620

139 Lomas, Natasha. 'How Facebook has reacted since the data misuse scandal broke', *Techcrunch* (2018). https://techcrunch.com/2018/04/10/how-facebook-has-reacted-since-the-data-misuse-scandal-broke/

140 Griffiths, J. 'Google fails to thread the china needle, again', *CNN* (2018). https://edition.cnn.com/2018/12/21/asia/google-china-dragonfly-analysis-intl/index.html

141 Shoham, Y., R. Perrault, E. Brynjolfsson, J. Clark, J. Manyika, J. C. Niebles, T. Lyons, J. Etchemendy, B. Grosz and Z. Bauer. 'The AI index 2018 annual report', *AI Index Steering Committee, Human-Centered AI Initiative*. Stanford University (2018).

142 Grace, K., J. Salvatier, A. Dafoe, B. Zhang and O. Evans. 'When will AI exceed human performance? Evidence from AI experts', *arXiv:1705.08807* (2017).

143 Dafoe, A. *AI Governance: A Research Agenda*. Governance of AI Program, University of Oxford (2018).

144 Brundage, M., S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel and H. Anderson. 'The malicious use of Artificial Intelligence: Forecasting, prevention, and mitigation', *arXiv:1802.07228* (2018).

145 Cave, S. and S. Ó hÉigeartaigh. 'An AI race for strategic advantage: Rhetoric and risks', *AAAI/ACM Conference on Artificial Intelligence, Ethics and Society* (2018).

146 Horowitz, M. C., G. C. Allen, E. B. Kania and P. Scharre. *Strategic Competition in an Era of Artificial Intelligence*. Center for a New American Security (2018); Allen, G. and T. Chan. *Artificial Intelligence and National Security*. Belfer Center for Science and International Affairs (2017); Rickli, J.-M. 'Artificial Intelligence and the future of warfare', *WEF Global Risks Report 2017* (2017), p. 49. http://www3.weforum.org/docs/GRR17_Report_web.pdf; Taddeo, M. and L. Floridi. 'Regulate artificial intelligence to avert cyber arms race', *Nature, 556*(7701) (2018): 296–98; Payne, K. *Strategy, Evolution, and War: From Apes to Artificial Intelligence*. Georgetown University Press (2018).

147 Geist, E. and A. J. Lohn. 'How might Artificial Intelligence affect the risk of nuclear war?', *RAND* (2018). https://www.rand.org/pubs/perspectives/PE296.html; Stoutland, P. O., S. Pitts-Kiefer, E. J. Moniz, S. Nunn and D. Browne. 'Nuclear weapons in the New Cyber Age: Report of the Cyber-Nuclear Weapons Study Group', *Nuclear Threat Initiative* (2018); Unal, B. and P. Lewis. *Cybersecurity of Nuclear Weapons Systems Threats, Vulnerabilities and Consequences*. Chatham House (2018); Lieber, K. A. and D. G. Press. 'The new era of counterforce: Technological change and the future of

nuclear deterrence', *International Security, 41*(4) (2017): 9–49; Kroenig, M. and B. Gopalaswamy. 'Will disruptive technology cause nuclear war?', *The Bulletin* (2018). https://thebulletin.org/2018/11/will-disruptive-technology-cause-nuclear-war/

148 Maas, M. M. 'Regulating for "normal AI accidents": Operational lessons for the responsible governance of artificial intelligence deployment', *Proceedings of 2018 AAAI/ACM Conference on AI, Ethics, and Society* (2–3 February 2018); Maas, M. M. 'How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons', *Contemporary Security Policy, 6* (2019): 1–27.

149 Pichai, S. *AI at Google: Our Principles* (2018). https://www.blog.google/technology/ai/ai-principles/; OpenAI. *OpenAI Charter* (2018). https://blog.openai.com/openai-charter/

150 OpenAI. *Better Language Models and Their Implications* (2019). https://blog.openai.com/better-language-models/

151 There is already a history of shareholder activism on the part of stock-holding employees of tech companies; for example, this year a group represented by Open MIC successfully tabled a shareholder resolution demanding that Amazon halt sales of its facial-recognition technology to law enforcement due to concerns over racial and gender bias — Open MIC. *A Win for Shareholders in Effort to Halt Sales of Amazon's Racially Biased Surveillance Tech* (press release) (2019). https://www.openmic.org/news/2019/4/4/a-win-for-shareholders-amazon

152 NASA. *NEOWISE* (2019). https://www.nasa.gov/mission_pages/neowise/mission/index.html

153 Baum, S. 'Risk-risk tradeoff analysis of nuclear explosives for asteroid deflection', *Risk Analysis, 39*(11) (2019): 2427–42.

154 Koplow, D. 'Exoatmospheric plowshares: Using a nuclear explosive device for planetary defense against an incoming asteroid', *UCLA Journal of International and Foreign Affairs, 1* (2019): 76–158.

155 Wee, Sue-Lee. 'China uses DNA to track its people, with the help of American expertise', *New York Times* (21 February 2019). https://www.nytimes.com/2019/02/21/business/china-xinjiang-uighur-dna-thermo-fisher.html

156 Slijper et al. (2019a); Slijper, F., Alice Beck, Daan Kayser and Maaike Beenes. 'Don't be evil? A survey of the tech sector's stance on lethal autonomous weapons', *PAX* (2019b); Slijper, F. 'The arms industry and increasingly autonomous weapons', *PAX* (2019).

157 Taken together the findings here suggest that institutional investors could consider global risk as a whole, rather than as single isolated hazards. If the motivation for investor campaigns is harm minimisation, the avoidance of risk, and continuation of an institution, investors should be concerned about reducing global risk in general rather than addressing specific threats. This would allow for investors to redirect finance from risks where investor campaigns are appropriate (climate, nuclear weapons and potentially LAWs and biotechnology) towards both specific mitigation strategies (renewable energy, asteroid deflection and pandemic preparedness) and general resilience to global risks (seed banks, disaster risk response technologies).

Contributors

SJ Beard is an Academic Programme Manager at the Centre for the Study of Existential Risk, University of Cambridge. https://orcid.org/0000-0002-2834-0993

Émile Torres is a a Postdoctoral researcher at the Inamori International Center for Ethics and Excellence, Case Western Reserve University. https://orcid.org/0000-0003-4420-9159

Luke Kemp is a Faculty Fellow at the Notre Dame Institute for Advanced Studies. https://orcid.org/0000-0002-7447-4335

Zoe Cremer is a Doctoral Student at the Human Information Processing Lab, University of Oxford.

Shahar Avin is a Senior Research Associate at the Centre for the Study of Existential Risk. https://scholar.google.com/citations?user=0-G2eiEAAAAJ

Bonnie C. Wintle is a Senior Research Fellow in the School of Agriculture, Food and Ecosystem Sciences, University of Melbourne. https://orcid.org/0000-0003-0236-6906

Julius Weitzdörfer is a Professor of Japanese Studies and East Asian Law at FernUniversität Hagen. https://scholar.google.com/citations?user=4Stozb8AAAAJ

Seán S. Ó hÉigeartaigh is theProgramme Director of AI: Futures and Responsibility at the Leverhulme Centre for the Future of Intelligence. https://orcid.org/0000-0002-2846-1576

William J. Sutherland is the Director of Research at the University of Cambridge Department of Zoology. https://orcid.org/0000-0002-6498-0437

Martin J. Rees is a Fellow of Trinity College and Emeritus Professor of Cosmology and Astrophysics at the University of Cambridge. He holds the honorary title of Astronomer Royal and also Visiting Professor at Imperial College London and at Leicester University. https://www.martinrees.uk/

Matthijs M. Maas is a Senior Research Fellow at the Institute for Law & AI. https://scholar.google.com/citations?user=Fe64DJQAAAAJ&hl=en

Hin-Yan Liu is an Associate Professor at the Centre for European, Comparative, and Constitutional Legal Studies, University of Copenhagen. https://scholar.google.com.my/citations?user=TO1tLV0AAAAJ&hl=en

Kristian Cedervall Lauta is Prorector for Education and Professor of disaster and climate law at the University of Copenhagen. https://scholar.google.com/citations?user=1DzaMFgAAAAJ

Adrian Currie is a Senior Lecturer in the Department of Sociology, Philosophy and Anthropology at Exeter. https://scholar.google.at/citations?user=ZAgjJKwAAAAJ&hl=en

Thomas Rowe is a Lecturer in Philosophy at King's College London. https://scholar.google.com/citations?user=RLIChuQAAAAJ

James Fox is a Doctoral Student at the Department of Computer Science, University of Oxford. https://scholar.google.com/citations?user=hMZs5tsAAAAJ

Mahlo N. C. Kennicutt is Professor Emeritus of Oceanography at Texas A&M University. https://scholar.google.com/citations?user=XpJYTbcAAAAJ

Gorm Shackleford is a former Research Associate at the Centre for the Study of Existential Risk. https://orcid.org/0000-0003-0949-0934

Rick Davies is an independent monitoring and evaluation consultant. https://richardjdavies.wordpress.com/

Lara Mani is a Senior Research Associate in Environmental Risk and Risk Communication at the Centre for the Study of Existential Risk, University of Cambridge. https://orcid.org/0000-0003-2967-7499

Tom Hobson is Postdoctoral Research Associate at the Centre for the Study of Existential Risk, University of Cambridge. https://orcid.org/0000-0002-1244-3787

Azaf Tzachor is an Associate Professor and the Academic Director of the Aviram Sustainability and Climate Program. https://orcid.org/0000-0002-4032-4996

Paul Cole is an Associate Professor of Volcanology at the School of Geography, Earth and Environmental Sciences, University of Plymouth. https://orcid.org/0000-0002-2964-311X

Catherine Richards is a Consultant working at McKinsey & Co. https://orcid.org/0000-0002-0084-0734

Richard Lupton is a Senior Lecturer in the Department of Mechanical Engineering, Centre for Sustainable and Circular Technologies (CSCT), University of Bath. https://orcid.org/0000-0001-8622-3085

Julian Allwood is a Professor of Engineering and the Environment at the University of Cambridge. https://orcid.org/0000-0003-0931-3831

Aaron Tang in a Doctoral Student at the Fenner School of Environment and Society, The Australian National University. https://orcid.org/0000-0002-3720-9831

Clarissa Rios Rojas is a Political Affairs Officer at UN Office for Disarmament Affairs. https://orcid.org/0000-0001-6544-4663

Catherine Rhodes is the Head of Operations and Engagement (SPRITE+)Head of Operations and Engagement (SPRITE+), University of Manchester. https://orcid.org/0000-0002-7747-2597

Paul Ingram is a Research Affiliate at the Centre for the Study of Existential Risk, University of Cambridge

Amritha Jayanthi is Deputy Chief Technology Officer at the U.S. Federal Trade Commission

Natalie Jones is a Policy Advisor at the International Institute for Sustainable Development. https://orcid.org/0000-0001-7060-0886

Ellen Quigley is a Principal Research Associate, the Co-Director of Finance for Environmental and Social Systemic Change, and the Special Adviser (Responsible Investment) to the Chief Financial Officer, all at the University of Cambridge. https://www.landecon.cam.ac.uk/person/dr-ellen-quigley

Index

About the Team

Alessandra Tosi was the managing editor for this book.

Rosalyn Sword copy-edited this book, and compiled the index.

Jeevanjot Kaur Nagpal designed the cover. The cover was produced in InDesign using the Fontin font.

Jeremy Bowman typeset the book in InDesign and produced the EPUB edition. The text font is Tex Gyre Pagella and the heading font is Californian FB.

Cecilia M. Thon produced the alt text.

Cameron Craig produced the PDF and HTML editions. The conversion is performed with open source software freely available on our GitHub page (https://github.com/OpenBookPublishers).

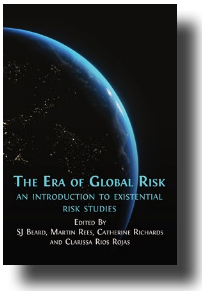
You may also be interested in:

The Era of Global Risk

An Introduction to Existential Risk Studies

SJ Beard, Martin Rees, Catherine Richards, Clarissa Rios Rojas (editors)

https://doi.org/10.11647/OBP.0336

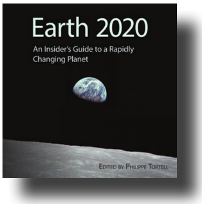

The Era of Global Risk
An Introduction to Existential Risk Studies
Edited By
SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas

Earth 2020

An Insider's Guide to a Rapidly Changing Planet

Philippe D. Tortell (editor)

https://doi.org/10.11647/OBP.0193

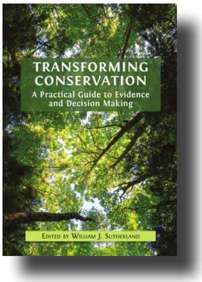

Earth 2020
An Insider's Guide to a Rapidly Changing Planet
Edited by Philippe Tortell

Transforming Conservation

A Practical Guide to Evidence and Decision Making

William J. Sutherland (editor)

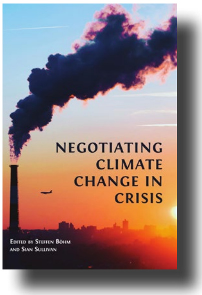
https://doi.org/10.11647/OBP.0321


Transforming Conservation
A Practical Guide to Evidence and Decision Making
Edited by William J. Sutherland

Negotiating Climate Change in Crisis

Steffen Böhm, Sian Sullivan (editor)

https://doi.org/10.11647/OBP.0265


Negotiating Climate Change in Crisis
Edited by Steffen Böhm and Sian Sullivan

An Anthology of Global Risk

Edited by SJ Beard and Tom Hobson

This anthology brings together a diversity of key texts in the emerging field of Existential Risk Studies. It serves to complement the previous volume *The Era of Global Risk: An Introduction to Existential Risk Studies* by providing open access to original research and insights in this rapidly evolving field. At its heart, this book highlights the ongoing development of new academic paradigms and theories of change that have emerged from a community of researchers in and around the Cambridge University Centre for the Study of Existential Risk. The chapters in this book challenge received notions of human extinction and civilization collapse and seek to chart new paths towards existential security and hope.

The interdisciplinary and trans-disciplinary nature of the cutting-edge research presented here makes this volume a key resource for researchers and academics. However, the editors have also prepared introductions and research highlights that will make it accessible to an interested general audience as well. Whatever their level of experience, the volume aims to challenge readers to take on board the extent of the multiple dangers currently faced by humanity, and to think critically and proactively about reducing global risk.

This is the author-approved edition of this Open Access title. As with all Open Book publications, this entire book is available to download for free on the publisher's website. Printed and digital editions, together with supplementary digital material, can also be found at http://www.openbookpublishers.com.

Cover image: Javier Miranda, Alien planet, June 18, 2022, https://unsplash.com/photos/nc1zsYG-kLFA
Cover design: Jeevanjot Kaur Nagpal

ebook
ebook and OA editions
also available

OPEN ACCESS

OpenBook
Publishers