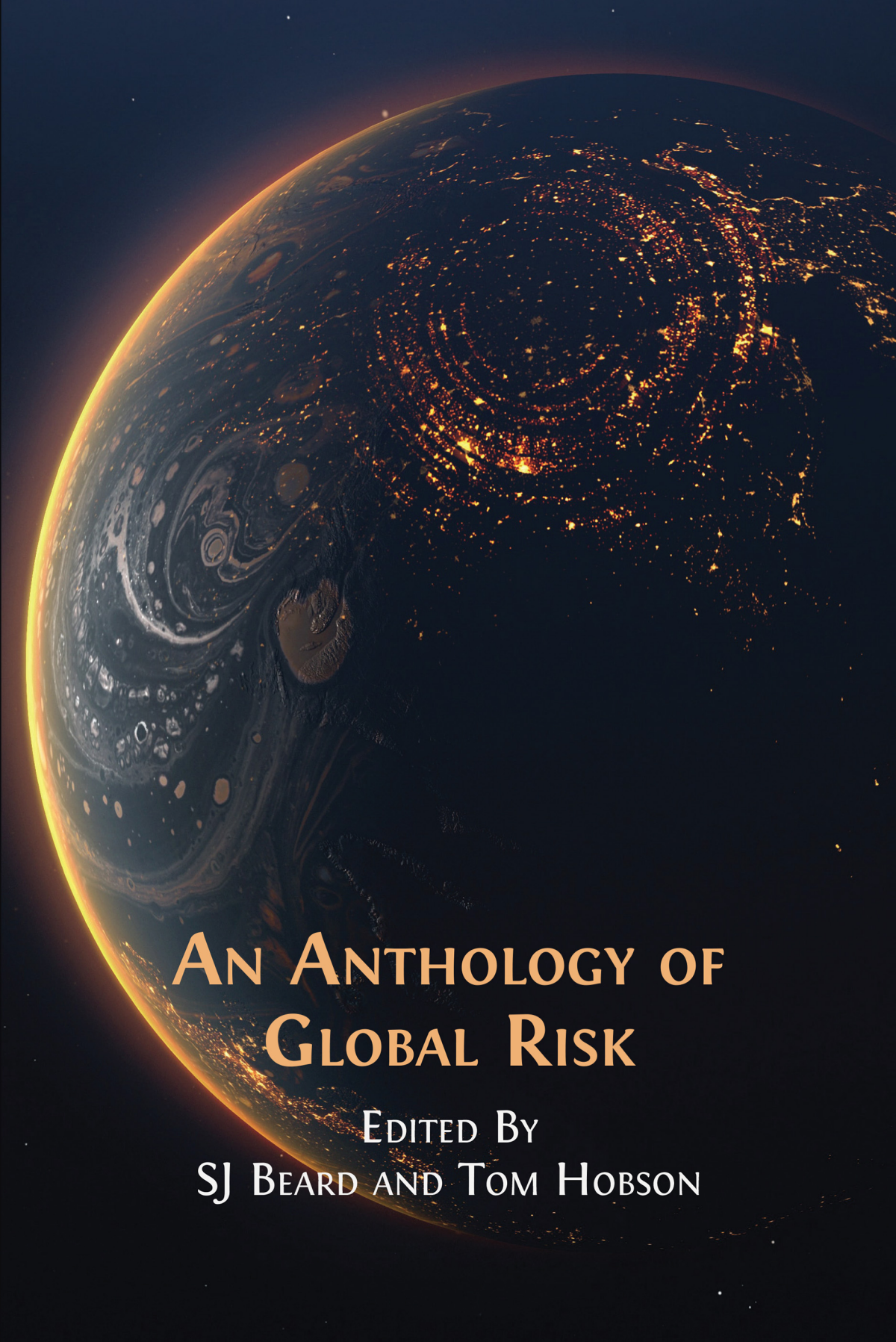

An Anthology of Global Risk

Edited By

SJ Beard and Tom Hobson

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

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.