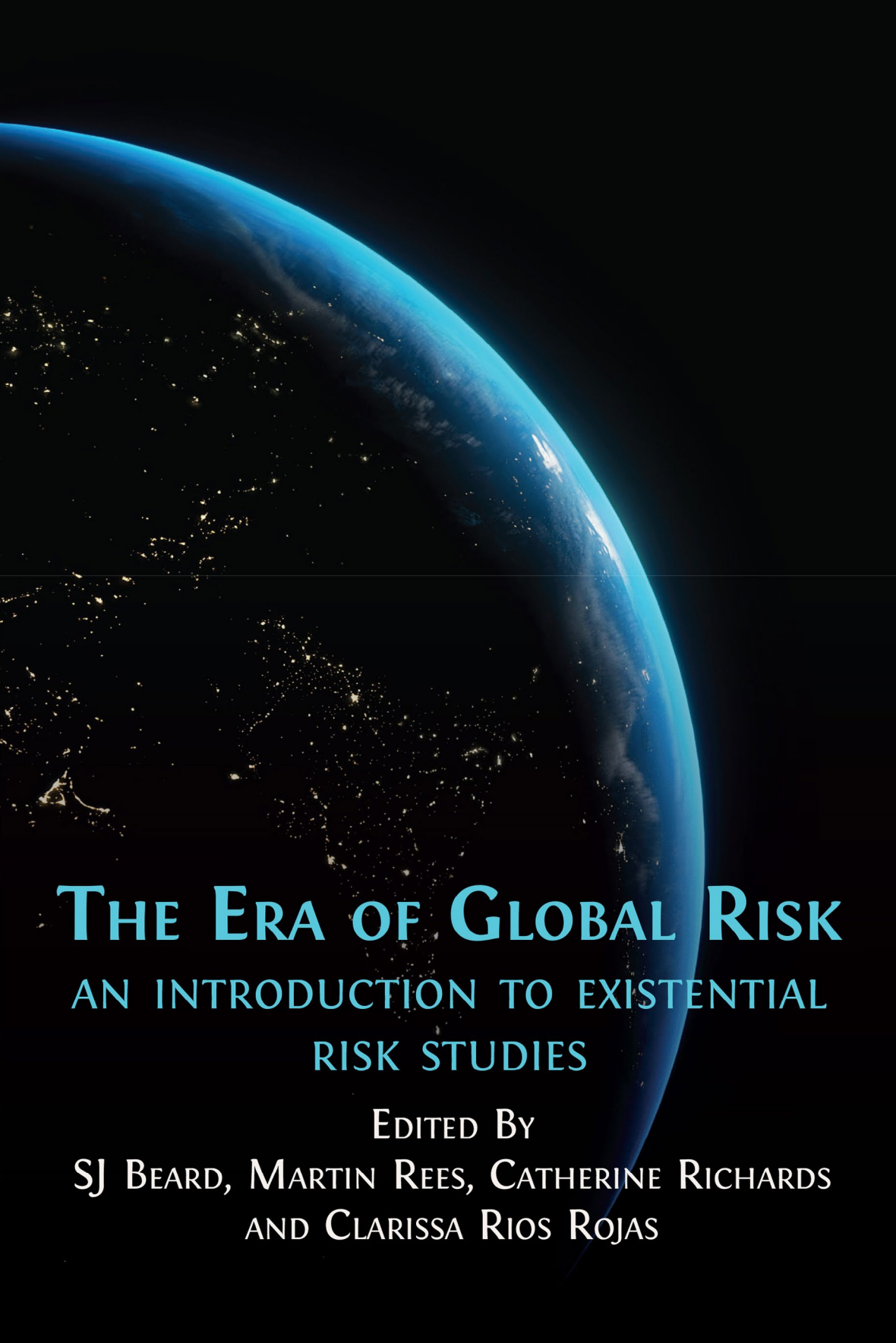

# The Era of Global Risk

## An Introduction to Existential Risk Studies

Edited By
SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas





SJ Beard, Martin Rees, Catherine Richards and Clarissa Rios Rojas (eds), *The Era of Global Risk: An Introduction to Existential Risk Studies*. Cambridge, UK: Open Book Publishers, 2023, https://doi.org/10.11647/OBP.0336



Digital material and resources associated with this volume are available at https://doi.org/10.11647/OBP.0336#resources

ISBN Paperback: 978-1-80064-786-2
ISBN Hardback: 978-1-80064-787-9
ISBN Digital (PDF): 978-1-80064-788-6
ISBN Digital ebook (epub): 978-1-80064-789-3
ISBN XML: 978-1-80064-791-6
ISBN HTML: 978-1-80064-792-3
DOI: 10.11647/OBP.0336

Cover image: Anirudh, *Our Planet* (October 14, 2021), https://unsplash.com/photos/Xu4Pz7GI9JY. Cover design by Jeevanjot Kaur Nagpal.

# Introduction

*SJ Beard, Martin Rees, Catherine Richards, and Clarissa Rios Rojas*

We are living in an era of global risk. While policymakers were once able to focus exclusively on the risks facing their particular constituency—be that a country, corporation, community, or institution—now, everybody must take account of the threats that endanger humanity as a whole. These come in many forms, from global-scale natural disasters (like volcanic super-eruptions) to anthropogenic environmental destabilisation (like climate change and loss of biosphere integrity), and from calamities that spread rapidly around our highly networked planet (like viruses and cyber threats) to the development of novel technologies with high destructive potential (such as artificial intelligence and biotechnologies). Reflecting this trend, the recent sixth edition of the United Nations Global Assessment Report on Disaster Risk Reduction, *Our World at Risk*, calls on member states for transformative governance that will lead to a resilient future, particularly given the increased occurrence and intensity of disasters. Similarly, the UN Secretary General's report, '*Our Common Agenda*', seeks to centralise the initiatives needed for better management to major global risks within discussions of global policy and governance.

One of the most prominent advocates for the importance of global risk has been the World Economic Forum, who defines a global risk as "the possibility of the occurrence of an event or condition that, if it occurs, could cause significant negative impact for several countries or industries". Since 2006, the Forum's annual *Global Risk Report*, based on a comprehensive risk perception survey of its members and stakeholders,

 https://doi.org/10.11647/OBP.0336.13

has provided something of a barometer showing which risks are of greatest concern. For instance, their inaugural report found that:

> The 2006 risk landscape is dominated by high impact headline risks, such as terrorism and an influenza pandemic, which top the global risk mitigation agenda and are increasingly well understood. Other risks, like climate change, whose cumulative impact will only be felt over the longer term, have begun to move to the centre of the policy debate and may offer the greatest challenges for global risk mitigation in the future.[1]

However, by the time of its most recent 2022 edition, the focus of the report has shifted markedly, now finding that over the next five years, leaders are most concerned about societal risks (such as social cohesion, livelihood, and mental health) and environmental risks, but that "over a 10-year horizon, the health of the planet dominates concerns: environmental risks are perceived to be the five most critical long-term threats to the world as well as the most potentially damaging to people and planet".[2] One trend that can be observed in this shift in risk perception is a long-term move away from concern about external threats we need to secure ourselves against (such as specific viruses or terrorism) and towards systemic risks that we, as human beings, are creating for ourselves, through poor governance, short-termism, and a too narrow focus on economic productivity.

Such a shift is very much in line with the developing understanding of global risk at the Centre for the Study of Existential Risk, located at the University of Cambridge. However, our concern is not simply to understand what risks decision-makers are most concerned by, but which ones they should be more concerned about, and what they need to do to mitigate those risks. There is an increasingly rich vocabulary for understanding global risks,[3] and with this, it has become clear that not all risks are the same. Some are also 'extreme', both in the sense that they involve extreme amounts of harm and that they could push global systems outside of their 'normal operating space'.[4] Within this class, two further subcategories have received particular attention. Global catastrophic risks (GCRs) involve events with one or more of the following characteristics: (a) "sudden, extraordinary, widespread disaster beyond the collective capability of national and international governments and the private sector to control";[5] (b) significant harm at the global scale, such as a large and sudden reduction in the global

population,[6] and/or (c) a failure of critical global systems,[7] including the cluster of sociotechnological systems we sometimes call 'human civilisation'. Finally, existential risks are those with the very worst potentialities, usually understood to involve either the extinction of humanity[8] or "the permanent and drastic destruction of its potential for desirable future development" (according to some assumptions about what desirable futures might be).[9] While these two are often conflated, it might be best to separate them into extinction risk and existential risk.[10]

There are many reasons why we should be especially concerned about extreme global risks, global catastrophic risks, and existential risks. Moral philosophers have argued that we have the strongest possible moral duty to mitigate these risks, whether on utilitarian,[11] idealist,[12] agent-centred,[13] or social-contract-based[14] grounds. Psychologists have also shown how people are systemically biased towards downplaying and ignoring these risks, and thus we need to work hard if we are to overcome these biases and give the risks the attention they deserve.[15] However, increasingly, we can also see that paying attention to risks such as these has tremendous practical importance. It seems likely that the current level of these risks is such that they could significantly impact the lives and futures of many people who are alive today, as well as being a significant threat to the long-term goals of many kinds of institution, from governments and charities to investors and corporations. In this book, we will not pay much attention to the reasons why one should focus on extreme global risks. Instead, we simply note that, if given the choice, most people would unquestionably want to protect themselves and others from such risks, and thus focus on the dual questions of how to understand these risks and manage them effectively.

The following ten chapters set out a number of different approaches to thinking about global, extreme, global catastrophic, and existential risks. The first five focus on the emerging science of global risk itself and build the case for an open and creative approach to studying these risks, drawing on lessons from the past—from the rich interdisciplinary literature on social and ecological collapse, from the experiences of people working on the governance of science and technology, from discussions about global injustice, and from the diversity of human beings with an interest in safeguarding our collective future. The second set of chapters then go on to provide more detailed assessments

of different risk drivers (including natural disasters, environmental breakdown, biotechnology, the potential of transformative future artificial intelligence (AI) in general, and the military application of AI in particular); the peculiar challenges to studying and mitigating each of these; and how they compare. Most, but not all, of these chapters were written by researchers affiliated to, or associated with, the Centre for the Study of Existential Risk at the University of Cambridge, and the chapters aim to provide those researchers' personal accounts of how best to think about this aspect of global risk while also engaging with, and surveying, a far broader range of literature and perspectives on the subject.

Our first chapter, 'A Brief History of Existential Risk and the People Who Worked to Mitigate It' by SJ Beard and Rachel Bronson, provides a historical account of our growing understanding of global risks and how scientists and others have worked to mitigate them. Looking back over the past 75 years, the chapter shows us how humanity has had to grapple with threats from nuclear weapons, environmental breakdown, and novel technologies to the political and technological forces that created them. However, it also surveys the many active scientific and political movements that have worked to avert disaster, as curious, compassionate, and courageous people have sought to understand these terrifying forces, bring them to wider public attention, and work to prevent human extinction and the collapse of civilisation. Using the iconic Doomsday Clock of the Bulletin of Atomic Scientists as a guide, it briefly tells the story of some of these people and organisations who sought to guide us safely through the 20$^{th}$ century and beyond. Understanding this history both helps us to understand the risks that continue to threaten humanity and offers opportunities to learn from the successes and failures of the past, rather than focusing only on whatever catastrophe is most immediate in our collective attention. In particular, the chapter highlights the importance of reinforcing key messages about risks, modelling extreme scenarios, managing the pace of scientific research, and placing its findings in the public domain—messages which are echoed in subsequent chapters.

The second chapter, 'Theories and Models: Understanding and Predicting Societal Collapse' by Sabin Roman, looks at what those who study global risks can learn from efforts to understand and model the

process of societal and ecological collapse, which is a significant global risk in itself and also an example of the kind of extreme, non-linear, and potentially dangerous transition that is associated with extreme global risks more generally. Surveying the extensive and interdisciplinary literature on this subject, in some cases extending back several centuries, the chapter illustrates the ways in which many qualitative and quantitative modelling approaches can be applied to shed insight on the causes and nature of such collapses. Some of these approaches are primarily concerned with the exogenous causes of collapse, such as conflict or environmental catastrophes. However, other approaches view collapse as endogenous to societies themselves, originating in economic inequality or shifting societal dynamics, and it is argued that even in the presence of external causes we cannot fully understand collapse unless we take account of these endogenous effects that ultimately make societies vulnerable in the first place. Perhaps most promisingly, the chapter indicates how we can create constructive new approaches based around modelling a variety of feedback loops between different elements, and how these can be adapted to generate and test new hypotheses about social and ecological collapse (either past or future).

Chapter 3, 'Existential Risk and Science Governance' by Lalitha S. Sundaram, looks at how the governance of science might matter for the production and prevention of existential risk, and whether there are options for making science and technology less risky that are being ignored. In particular, it focuses on the ways in which scientific governance is conventionally framed within the global risk community—as something extrinsic to be regulated, with either greater top-down control to promote safety, or greater libertarian freedom to promote innovation—and highlights the potential shortcomings of this approach. As an alternative, it proposes considering scientific governance more broadly as a constellation of socio-technical processes that shape and steer technology, and argues that research culture and self-governance within science need to be seen as central to how science and technology developments play out. This alternative framing highlights many new levers at our disposal for ensuring the safe and beneficial development of technologies; overlooking such possibilities could mean robbing humanity of some of our most effective tools for mitigating global risk. The chapter ends by proposing some areas where scientists and the

global risk community might together hope to influence those existing modalities, such as via education, professional bodies, two-way policy engagement, collective action, and public outreach.

Chapter 4, 'Beyond "Error and Terror": Global Justice and Global Catastrophic Risk' by Natalie Jones, serves as an invitation to consider global political, economic, social, and legal systems (particularly in relation to global justice and inequality) when studying and addressing global catastrophic risks. While the previous chapter showed how our understanding of risk was hampered by too great a focus on top-down approaches to mitigation and governance, this chapter highlights a no less important blind spot in much of the thinking about global risk: the tendency to focus more on individuals and institutions as agents of risk, and neglect the importance of systems of extraction, oppression, marginalisation, and corruption. While individuals and institutions are undoubtedly important drivers of global risk, studying global risk while ignoring global injustice can distort our understanding of risk. In contrast, adding a global justice lens onto our existing strategies helps us see the nature of risks more clearly. Furthermore, as the case of climate change shows, strategies to reduce global catastrophic risk will be more effective if they take account of global justice considerations. It follows that policies to reduce global catastrophic risk can—and should—be designed to simultaneously mitigate risk and achieve justice.

Chapter 5, 'We Have to Include Everyone: Enabling Humanity to Reduce Existential Risk' by Sheri Wells-Jensen and SJ Beard, argues for the importance of considering diversity and inclusion as integral both to a flourishing science of global risk and to efforts to mitigate such risks. Given the scale and importance of global risk, it can be tempting to believe that only the most able would be able to understand and mitigate it effectively. However, the chapter argues that such thinking is clearly mistaken. Far from being merely vulnerable and unable to help, disabled people and others who are marginalised or excluded are the real experts in vulnerability, adaptation, and resilience, and have a lot to contribute to studying and managing risks, even on the global scale. Moreover, diversity and inclusion are vital sources of creativity and insight. This chapter explores the limitations and costs of standard narratives around diversity and inclusion in global risk, and shows how the global risk community would benefit from championing inclusive futures and

paying more attention to disabled people and other marginalised groups. It focuses on the benefits of diversity and inclusion across three case studies (foresight and horizon scanning, space colonisation, and bioethics) to highlight this point, while also considering the wider costs of marginalisation and exclusion to society as a whole.

Moving onto specific drivers of risk, Chapter 6, 'Natural Global Catastrophic Risks' by Lara Mani, Doug Erwin, and Lindley Johnson, considers risks from 'natural' disasters. It explores the dichotomies that are often neglected and left on the peripheries of discussions about such risks falling somewhere between hazard and vulnerability. The chapter shares a similar perspective to Chapter 2, that while the historical and geological record of such disasters can be used to study their impact, we need to consider more than just the rate of disasters as exogenous events and also take account of the factors that make societies and species more or less vulnerable to them if we are to understand the evolving nature of this risk. The chapter argues that, while humanity has lived with global-scale natural threats (such as large magnitude volcanic eruptions and Near-Earth Object impacts) throughout history, the risk of such events is currently growing due to the increasing scale and complexity of human society. Thus, while the probability of potentially catastrophic natural hazards of this kind may be relatively low, it is certainly not negligible, and the societal and economic impacts are potentially vast; however, this type of hazard is frequently underestimated in the literature. The chapter surveys the state of current thinking around extreme natural risks and asks what can be learned from efforts to reduce some of these risks (such as Planetary Defense against near-Earth objects) for improving our resilience to natural global-scale catastrophes more generally.

Chapter 7, 'Ecological Breakdown and Human Extinction' by Luke Kemp, explores the catastrophic potential of anthropogenic environmental risks, and (in particular) climate change. The chapter considers both the scale and nature of global risk from climate change and the arguments for prioritising climate mitigation as a way of reducing global risk. Reviewing the available evidence, it notes the weaknesses of certain arguments that climate change is and is not a risk with global catastrophic and existential potential. However, while there are many plausible reasons to be concerned about the catastrophic potential of climate change, it finds that attempts to argue that we

should not consider climate change as being of the same severity as technological global risks often depend upon spurious notions of what a climate-induced catastrophe might involve. It then considers the appropriateness of using existing discourse around existential and global catastrophic risk to talk about climate change in the first place, given that this often frames risks in terms of their potential impact on long-term economic and technological growth, which is a questionable goal and one that (in many ways) assumes that possible ecological limits to human growth should be disregarded out of hand. Finally, however, in considering the case for climate mitigation as a global risk reduction strategy, the chapter makes the case that there is compelling evidence in favour of this, not only due to the direct impacts of climate mitigation but also the substantial co-benefits to human health and flourishing that many policies aimed at climate mitigation might provide. However, it also argues that many of the strategies proposed for climate mitigation at the global scale are problematic because they misidentify the root causes of the problem in identifying climate change as a 'tragedy of the commons' when it is actually a 'tragedy of the elite' where, as previously discussed in Chapter 4, systems of global injustice are empowering a small number of agents with the capacity to do large amounts of harm and also incentivising them to do so.

Chapter 8, 'Biosecurity, Biosafety, and Dual Use: Will Humanity Minimise Potential Harms in the Age of Biotechnology?' by Kelsey Lane Warmbrod, Kobi Leins, and Nancy Connell, discusses a number of recent advances in the life sciences that may serve to contribute to the current level of global risk (both positively and negatively), their convergence with developments in many other fields (such as AI and nanotech), and the harms that might be caused by their misuse. The chapter surveys recent developments across genomics, gain of function experiments, gene drives, synthetic biology, and AI-enabled biological research. It contrasts the rapid development and interdisciplinarity of these fields with the slow-moving pace of efforts to govern their use, often relying on the now decades-old Biological Weapons Convention. It thus emphasises the need for new approaches that fully embrace the power and flexibility of bottom-up science governance, as described in Chapter 3, and also the empowerment of communities who are often disproportionately affected by the quest for new technologies, as advocated for in Chapters 4 and 5. Grappling with the multiple

potentialities of new technologies requires careful thought, but it also requires researchers and practitioners to work collectively to address the challenges we currently face as biology marches towards a global bioeconomy. This is an achievable goal but will require action to be taken soon. Urgent actions include creating and conducting a robust risk assessment methodology and implementing appropriate biosafety measures; strengthening frameworks for obtaining and enforcing consent for research, including at the community level; and requiring higher standards of interpretability for algorithms and big datasets used in biological research and the development of biotechnologies (a problem also discussed in the next chapter).

Chapter 9, 'From Turing's Speculations to an Academic Discipline: A History of AI Existential Safety' by John Burden, Sam Clarke, and Jess Whittlestone describes the development of thought related to artificial intelligence (AI) and existential risk. These risks are more likely to be realised by future AI systems with greater capabilities and generality than current systems; however, the field of AI is moving extremely swiftly and AI systems are becoming more ubiquitous in the daily lives of people around the world. Great care must, therefore, be taken to ensure these systems are safe. The chapter describes how the field of existential AI safety has matured from pure speculative concerns in the 20$^{th}$ century into a rigorous academic discipline of technical expertise. In particular, it focuses on the problem of *alignment*. An AI system is considered aligned if it behaves according to the values of a particular entity, such as a person, an institution, or humanity as a whole. There are many ways in which AI systems may become misaligned, or in which the need for different alignments may pull it in conflicting directions, and the problem could thus arise in a wide variety of contexts, with different but no less serious existential consequences in each of these. Just as important as our evolving understanding of the problems of AI safety, however, have been the development of new approaches to achieving AI safety and ensuring meaningful—and beneficial—human control over AI systems. Furthermore, despite the significant progress that has been made, the field remains surprisingly small, and its recent history only serves to highlight the many prospects for further development in the near future.

Finally, Chapter 10, 'Military Artificial Intelligence as a Contributor to Global Catastrophic Risk' by Matthijs M. Maas, Kayla Lucero-Matteucci,

and Di Cooke, focuses specifically on the uses of AI to increase humanity's destructive capabilities within the military context. After reviewing past military GCR research and recent pertinent advancements in military AI, the chapter focuses on lethal autonomous weapons systems (LAWS) and the intersection between AI and nuclear weapons, both of which have received the most attention thus far. Regarding LAWS, it argues that, while the destructive capabilities of this technology are increasing, it is unlikely these will constitute a global catastrophic or existential risk in the near future, based primarily on current and anticipated costs and production trajectories. On the other hand, it argues that the application of AI to nuclear weapons has a significantly higher GCR potential. The chapter cites the danger of this within existing debates over when, where, and why nuclear weapons could lead to a GCR, as well as the recent geopolitical context, by identifying relevant converging global trends that may be raising the risks of nuclear warfare. The chapter turns its focus to the existing research on specific risks arising at the intersection of nuclear weapons and AI, and outlines six hypothetical areas where the use of AI systems in, around, or against nuclear weapons could increase the likelihood of nuclear escalation and result in global catastrophes. These systems include the automation of nuclear decision-making, the pressurisation of human decision-making, AI deployment in systems peripheral to nuclear weapons, AI as a threat to information security, AI as a threat to nuclear integrity, and broader impacts on strategic stability. The chapter concludes with suggestions for future directions of study, and sets the stage for a research agenda that can gain a more comprehensive and multidisciplinary understanding of the potential risks from military AI, both today and in the future.

While nowhere near fully comprehensive in scope, these chapters provide a snapshot of a rapidly evolving field: the scientific study of global risk as a phenomenon of urgent but tractable problems with global importance.[16] It is a field that, although undergoing significant growth in recent years, still remains surprisingly small and neglected. Unfortunately, it is also a field that is already showing signs of disciplinary fracture (for instance, between researchers working primarily on environmental risks and those working primarily on technological risks) that desperately needs to be understood and addressed. This book represents the first interdisciplinary survey of

the topic to come out since Nick Bostrom and Milan Cirkovic's *Global Catastrophic Risks* in 2008,[17] and its intention is precisely to provide both a survey and prospectus for this science as a vibrant, open, and rigorous field of academic research. Each of these chapters presents a clear call for action and has been specially written with an educated lay audience in mind, although we submit that, given the range and nature of material being presented, they may not always be for the faint of heart. Nevertheless, we believe that, in this era where no one can ignore the threats that endanger all humanity, it is imperative that this science should be available to all, and that everyone should ask themselves: what is my role and how can I contribute to bringing the era of global risk to a close and move towards an era of global safety?

## Acknowledgements

The editors of this volume would like to thank Clare Arnstein, Esmé Booth, Laura Elmer, Alice Jondorf, Jess Bland, Seán Ó hÉigeartaigh, and the staff at Open Book Publishing for invaluable assistance in preparing this volume. Many of the chapters here grew out of panel discussions from the 2020 Cambridge Conference on Catastrophic Risk at the Centre for the Study of Existential Risk and we would also like to thank Lara Mani, Catherine Rhodes, and Annie Bacon for helping to organise this as well as the other panellists and participants for their contributions. This publication was made possible through the support of 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 World Economic Forum, *Global Risks 2006* (2006).

2 World Economic Forum, *The Global Risks Report 2022: 17th Edition* (2022).

3 Cremer, Carla Zoe and Luke Kemp, 'Democratising risk: In search of a methodology to study existential risk', *arXiv preprint arXiv:2201.11214* (2021); Sundaram, Lalitha S., Matthijs M. Maas and S.J. Beard, 'From Evaluation to Action: Ethics, Epistemology and Extreme Technological Risk' in Catherine Rhodes (ed), *Managing Extreme Technological Risk. World Scientific Publishing* (forthcoming).

4 Broska, Lisa Hanna, Witold-Roger Poganietz, and Stefan Vögele, 'Extreme events defined—A conceptual discussion applying a complex systems approach', *Futures, 115* (2020), p.102490.

5 Schoch-Spana, Monica, Anita Cicero, Amesh Adalja, Gigi Gronvall, Tara Kirk Sell, Diane Meyer, Jennifer B. Nuzzo, et al. 'Global catastrophic biological risks: Toward a working definition', *Health Security, 15*(4) (2017), pp.323-328.

6 For instance: Cotton-Barratt Owen, Sebastian Farquhar, John Halstead, Stefan Schubert, and Andrew Snyder-Beattie, *Global Catastrophic Risks 2016.* Global Challenges Foundation (2016) use a 10% reduction; Kemp, Luke, Chi Xu, Joanna Depledge, Kristie L. Ebi, Goodwin Gibbins, Timothy A. Kohler, Johan Rockström et al. 'Climate Endgame: Exploring catastrophic climate change scenarios', *Proceedings of the National Academy of Sciences, 119*(34) (2022): e218146119 use a 25% reduction (while referring to risks involving a 10% reduction as 'Decimation Risks'), and Maas, Lucero-Matteucci, and Cooke (this volume) use a threshold of 1 million fatalities.

7 Avin, Shahar, Bonnie C. Wintle, Julius Weitzdörfer, Seán S. Ó hÉigeartaigh, William J. Sutherland, and Martin J. Rees, 'Classifying global catastrophic risk', *Futures, 102* (2018), pp.20-26.

8 Kemp et al. (2022).

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

10 Cremer and Kemp (2021).

11 Bostrom, Nick, 'Existential risk prevention as global priority,', *Global Policy, 4*(1) (2013), pp.15–31.

12 Parfit, Derek, *Reasons and Persons*. Oxford University Press (1984); Beard, S.J. and Patrick Kaczmarek, 'On Theory X and what matters most', *Ethics and Existence: The Legacy of Derek Parfit* (2021), p.358.

13 Scheffler, Samuel, *Why Worry About Future Generations?* Oxford University Press (2018).

14 Finneron-Burns, Elizabeth, 'What's wrong with human extinction?', *Canadian Journal of Philosophy, 47*(2–3) (2017), pp.327–43; Beard, S.J. and Patrick Kaczmarek, 'On the wrongness of human extinction', *Argumenta, 5* (2019), pp.85–97.

15 Yudkowsky, Eliezer, 'Cognitive biases potentially affecting judgement of global risks', *Global Catastrophic Risks, 1*(86) (2008), p.13.

16 One work that is foundational for both this science in general and this volume in particular was published by Martin Rees in 2003. It was originally intended that this work should be titled *Our Final Century?*; however, its publishers sought to outdo one another in making this sound more alarmist, first removing the question mark for the UK edition and then substituting 'Hour' for 'Century' for the American market. For this reason, the international group of authors behind these chapters cite this important work as both Rees, M., *Our Final Century: Will Civilisation Survive the Twenty-First Century?* Random House (2003) and Rees, M., *Our Final Hour: A Scientist's Warning*. Basic Books (2003).

17 Bostrom, Nick and Milan M. Ćirković (eds), *Global Catastrophic Risks*. OUP (2008).