# Identifying and Assessing the Drivers of Global Catastrophic Risk: A Review and Proposal for the Global Challenges Foundation

*Simon Beard and Phil Torres*

**Summary**

The goal of this report is to review and assess methods and approaches for assessing the drivers of global catastrophic risk. The review contains five sections:

(1) A conceptual overview setting out our understanding of the concept of *global catastrophic risks* (GCRs), their drivers, and how they can be assessed.
(3) A brief historical overview of the field of GCR research, indicating how our understanding of the drivers of GCRs has developed.
(2) A summary of existing studies that seek to quantify the drivers of GCR by assessing the likelihood that different causes will precipitate a global catastrophe.
(4) A critical evaluation of the usefulness of this research given the conceptual framework outlined in section 1 and a review of emerging conceptual, evaluative and risk assessment tools that may allow for better assessments of the drivers of GCRs in the future.
(5) A proposal for how the Global Challenges Foundation could work to most productively improve our understanding of the drivers of GCRs given the findings of sections 2, 3, and 4.

Further information on these topics will be included in three appendices, each of which is an as-yet-unpublished academic paper by the authors, alone or in collaboration with others. These are:

(A) Transhumanism, Effective Altruism, and Systems Theory: A History and Analysis of Existential Risk Studies.
(B) An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards.
(C) Assessing Climate Change's Contribution to Global Catastrophic Risk.

This report highlights how the body of emerging GCR research has failed to produce sufficient progress towards establishing a unified methodological framework for studying these risks. Key steps that will help to produce such a framework include moving away from a hazard-focused conception of risk, typified by the majority of quantitative risk assessments that we analyze, and toward a more sophisticated approach built on a mature understanding of risk assessment and disaster risk reduction and preparedness. We further suggest that a key barrier to the development of a mature science capable of comprehensively assessing the drivers of GCRs has been the political, philosophical, and economic context within which the field has arisen, as demonstrated by five distinct "waves" of GCR research. We believe that a suitably committed funder that transcends these contextual boundaries could have a transformative impact on the discipline, and with it our understanding of GCRs and their drivers. We propose that the Global Challenges Foundation is in a uniquely strong position to play this role.

**Contents**

## Section 1: Conceptual Issues

### 1.1 Global Catastrophes and their Risks

When people analyze and study the drivers of global catastrophic risks (GCRs), it is typical to do so in terms of two key features of risk, namely, *scale* and *severity*.[1] Thus a GCR may be defined as having "the potential to inflict serious damage to human well-being on a global scale";[2] risks that cause "significant harm" to "the entire human population or a large part thereof";[3] "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";[4] and "events [which] could lead to sudden, extraordinary, widespread disaster beyond the collective capability of national and international governments and the private sector to control."[5]

We believe that this represents an impoverished framework for risk classification and assessment. To begin with, definitions like those above do not describe *risks* so much as *catastrophes*. None mention key risk concepts like probability, ineluctability, and mitigation. As we examine in a later section of this report, these problems likely reflect the overarching concerns of the researchers who first engaged with the general topic of GCRs (understood as the subject of a unified field of research) and who approached the topic from a particular techno-utopian and transhumanist perspective rather than from the perspective of risk management and analysis. The former defines a set of trajectories for the future of humanity, some of which could instantiate an "existential risk" (a subtype of GCR), which Nick Bostrom defines essentially as a failure to attain a stable state of "technological maturity" (i.e., the total subjugation of nature and maximization of economic productivity). We believe that this perspective is limited both in its research agenda and the range of risks it identifies as important.

A second weakness in the above GCR definitions is that while they seek to present a unified concept of what a GCR is, they do not present a unified view of GCR *itself*. This is because in defining something "as" a GCR, they tend to conflate the threats or hazards that could bring about a catastrophe and the *risk* of the catastrophe. This is problematic. First, doing so fails to notice that we face many interconnected risks and that one risk driver might serve both as a direct hazard to humanity and as a multiplier or intensifier of other risk drivers. Thus, while many scholars have debated whether or not climate change is a GCR (on some definition), the fact is that this phenomenon is a major *contributor* to GCRs, although its contribution is often mediated through other risk drivers, such as food insecurity, international conflict, loss of biosphere integrity, and future geoengineering technologies, none of which can be fully understood in isolation from the others. Second, by focusing on the hazards that drive risk (i.e., the catastrophic events or processes

[1] This framework originated in Bostrom, N., 2002. Existential risks: Analyzing human extinction scenarios and related hazards. *Journal of Transhumanism*, 9. The structure and diversity of definitions of the related concepts of "global catastrophic risk" and "existential risks" are similar and many authors treat the two terms as derivatives of one another. See also, Torres, P., 2019. Existential Risks: A Philosophical Analysis. *Inquiry*, pp.1-26.
[2] Bostrom, N. and Cirkovic, M.M. eds., 2008. *Global Catastrophic Risks*. OUP Oxford.
[3] Bostrom, N., 2013. Existential Risk Prevention as Global Priority. *Global Policy*, 4(1), pp.15-31.
[4] Cotton-Barratt, O., Farquhar, S., Halstead, J., Schubert, S. and Snyder-Beattie, A., 2016. Global Catastrophic Risks 2016. Global Challenges Foundation.
[5] Schoch-Spana, M., Cicero, A., Adalja, A., Gronvall, G., Kirk Sell, T., Meyer, D., Nuzzo, J.B., Ravi, S., Shearer, M.P., Toner, E. and Watson, C., 2017. Global Catastrophic Biological Risks: Toward a Working Definition. *Health Security*, 15(4), pp.323-328.

themselves), it ignores the crucial questions of why and how humanity is *vulnerable* to such hazards and *exposed* to such harm. Thus, despite their popularity within the GCR community, we will not attempt to incorporate any of these concepts of GCR directly into this report.

Instead, we begin with the observation that at the global level, catastrophic risk represents the failure of governance strategies for preventing global harm. While individuals may strongly differ about what they would like humanity's future to be, any reasonable global governance framework would see certain outcomes as intrinsically unacceptable. These are outcomes that imply either a permanent end to humanity or at least a cessation of our ability to collectively provide what John Rawls refers to as the "basic goods" necessary for human flourishing.[6] At the Centre for the Study of Existential Risk (CSER), we describe these as human extinction and civilization collapse, respectively, although it is worth nothing that these terms have definitional issues of their own are meant merely as ordinary language glosses on a more precisely determined undesirable end-state that would permanently curtail humanity's future potential.

Given this, we can understand GCRs as global governance failures and/or fragilities. These should be understood in probabilistic terms, but should also be seen as ultimately under human control, representing vulnerabilities, exposures, or hazards that we might prevent or manage but have not. The possibility of such failures is, by its very nature, a governance failure: an inability for humanity to collectively achieve our basic common ends. This is important for assessing the drivers of GCRs at a number of levels.

Firstly, it fundamentally links GCRs not only with risk, i.e., the chance of reaching some undesirable outcome, but with the need to avoid this risk. It therefore situates GCRs at the level of global governance and sees it as an endogenous feature of human civilization. Early scholars of GCRs tended to be inspired by exogenous catastrophes like asteroid strikes and volcanic supereruptions and thus theorized about anthropogenic GCRs as if they could be studied as isolated instances of human failure. Thus, there have been several studies that seek to quantify the likelihood of global catastrophe without recognizing that such estimates are, in essence, judgements about our collective ability to understand and respond to our own failings.[7] This has also led several scholars to claim that certain drivers of risk, most notably from environmental factors such as climate change, can be ignored because of the potential to devise technological fixes, without an apparent understanding that it is the very achievability of such fixes that determines the level of this risk.

A second feature of this understanding of GCRs that needs highlighting is that it immediately focuses attention onto the management and prevention of risks, which is where it should rightly be. It is of very little value postulating about whether the level of existential risk facing humanity over the next century is 50%,[8] 19%,[9] or 1%[10] (or even whether AI contributes more to this level of risk than climate change), unless this helps to understand and address the underlying causes of the GCRs facing humanity. It would be better to formulate a robust plan for lowering the level of GCR

[6] Rawls, J., 1999. *A Theory of Justice*. Harvard University Press.
[7] We highlight a number of these in section 2 below.
[8] Rees, M.J., 2003. *Our Final Century*. Basic Books.
[9] Sandberg, A. and Bostrom, N., 2008. Global Catastrophic Risks Survey. Future of Humanity Institute, Technical Report. https://www.fhi.ox.ac.uk/reports/2008-1.pdf.
[10] Metaculus online prediction market, 2019. Will Humans Go Extinct by 2100? https://www.metaculus.com/questions/578/human-extinction-by-2100/

while not knowing how serious our situation is than to understand how serious our situation is without knowing how to lower the level of GCR.

1.2 The Foundations of Global Catastrophic Risk Assessment

A sound conception of GCRs is not, however, sufficient to assess the levels and drivers of this risk. To this conception of what GCRs are and how it should be addressed we thus intend to add additional insights from the fields of risk analysis and disaster studies that were not well incorporated by previous GCR research.

One of these, which we will describe in detail below, is that risks are created by the convergence of hazards, vulnerabilities, and exposures. Put simply, a *hazard* is some event that might affect a critical system on which humanity (or anything else we choose to care about) depends. A *vulnerability* is any feature of human systems that means that the hazard will have this effect. An *exposure* is the feature of this effect, arising from the convergence of the hazard and vulnerability, that makes it harmful.

This more sophisticated classification of risks is important for several reasons. The first of these is that when people think of the drivers of GCRs they predominantly consider the hazards currently facing humanity, such as asteroids, AI, pandemics, climate change, and so on. However, an equal if not greater contributor to the current level of GCRs may be society's lack of resilience, a combination of (a) our reliance on numerous ecological, sociotechnological, economic, and cultural systems to operate within quite narrowly defined operating parameters, currently with very little redundancy or flexibility (a clear vulnerability), and (b) our susceptibility to damages resulting from a range of factors as diverse as pathogens and weapons to a loss of social connectedness and ecosystem services (an exposure). We often fail to recognize this lack of resilience because we have increasingly come to optimize our societies to function within the limited parameters that our industrialized global economy has experienced over the course of its history—at most 500 years. However, when considered against the wider sweep of conditions that humanity, not to mention other species, have experienced during geological time, and the very rapid and interlinked changes currently affecting this multiplicity of systems, it is not difficult to see the many vulnerabilities and exposures that exist and that, in the absence of many hazards over the past 50 years, have been growing via the pursuit of economic efficiency and political expediency.

Unfortunately, while we are convinced that this is the best way to conceptualize and assess GCRs, the vast bulk of the literature in this field has taken a far more limited conception as their starting point. Rather than dismissing this important literature, however, our goal here will be to begin by summarizing it and then critically evaluating the assumptions behind it. We can then understand how such existing work can be repurposed towards the goal of a fuller, more sophisticated, governance-focused study of the drivers of GCRs before setting out our proposals for how we can most easily move towards a richer and more useful understanding of this topic—a research agenda in which we believe and hope that the Global Challenges Foundation can play a leading role.

## Section 2: How the Field of Global Catastrophic Risk Research Has Developed

People in all cultures throughout history have speculated about the possibility of global catastrophes, up to and including the "end of the world" or "apocalypse." However, these

speculations have largely been bound up with religious beliefs.[11] In contrast, the notion of global catastrophic risk (GCR) is *naturalistic*: it does not reference any supernatural entities or phenomena, which thus renders the concept appropriate for scientific investigation. Here we identify five historical "waves" of thinking about GCRs and related issues. These are differentiated by their key concepts, ethical assumptions, and methods, the scope of their inquiries, and their various intellectual constraints. This section offers a brief summary of each wave; in section 4, we indicate how developments in the final wave, inspired by global systems theory, are opening up new approaches to understanding GCRs and their drivers.

2.1 First Wave: Speculative Fiction

Some of the earliest thoughts about human extinction in a naturalistic sense are found in the writings of literary notables like Lord Byron, Mary Shelly, Camille Flammarion, and H.G. Wells. Lord Byron, for example, is reported to have believed that humanity would someday perish as a result of a comet impact, and that this has happened many times before on Earth. In his poem "Darkness," he imagines a future in which Earth becomes lifeless.[12] Ten years later, Shelly published *The Last Man*, which tells the story of Lionel in the last few decades of the twenty-first century. After a series of apocalyptic events, most notably a worldwide plague, Lionel becomes the only remaining survivor. This novel exemplifies the theme of the "last man," which had become popular by the 1820s following the first novel of the genre (and title) by Jean-Baptiste Cousin de Grainville, in 1805, which speculated about a future in which the human population dwindles because humans become incapable of procreating.

The discovery of the second law of thermodynamics in the 1860s inspired new thoughts about human extinction among both science fiction writers and working scientists. For example, in his 1870 book *Sketches of Creation*, the American geologist Alexander Winchell "describes the 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."[13] Similarly, *The Time Machine* by H.G. Wells tracks the adventures of an anonymous time-traveler who ventures to the end of the world, which he finds cold and nearly lifeless. This is long after humanity has gone extinct due to the sun burning out. Other writers considered the future of humanity from an evolutionary perspective, following Charles Darwin's 1859 *Origin of Species*. For example, in *First and Last Men* (1930), 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 thirty-five people who split into two groups. Although our evolutionary lineage persists, *Homo sapiens* dies out.

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 Shelly's *Frankenstein* (1818). The first novel to mention a technological accident destroying the world may have been Jules Verne's 1862 *Five Weeks in a Balloon*. In it, one character states: "I sometimes think that the end of the

[11] Torres, P., 2016. *The End: What Science and Religion Tell Us About the Apocalypse*. Pitchstone Publishing; and Moynihan, T., 2019. Existential Risk and Human Extinction: An Intellectual History. *Futures*, 116.
[12] Byron, G., 1815. Darkness. https://poets.org/poem/darkness.
[13] Shields, C.W. 1877. *The Final Philosophy*. Scribner, Armstrong, and Co.

world will come when some immense boiler, heated to three thousand atmospheres, blows up the earth." 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. Stanley Kubrick's 1964 film *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb* exemplified the worries shared by many during the Cold War of a nuclear holocaust. The film ends with a Soviet Union "doomsday machine" spreading lethal radiation around the world.

Writers of speculative fiction where 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 developing machines to hurl rocks into the sky in order to knock oncoming comets from its course.[14] Similarly, William Hope Hodgson's *The Night Land* (1912) depicts the plight of humanity after the sun has burned out, but describes our species surviving in huge pyramids that are geothermally heated, with crops grown underground in hydroponic rooms, while the 1923 novel *Nordenholt's Million*, by Alfred Walter Stewart, tells the story of a plutocrat who creates a refuge in Scotland after an engineered "denitrifying" bacteria causes the food supply to collapse.

Central themes of this wave include the "last man," the inevitability of disaster, the folly of hubris, and the secular discovery of the future. According to W. Warren Wagar, Wells founded the field of "future studies" with his 1901 book *Anticipations of the Reaction of Mechanical and Scientific Progress upon Human Life and Thought*,[15] followed by his 1902 Royal Institute lecture titled "The Discovery of the Future," although Amual Butler's essay "Darwin Among the Machine's" (1863) and his consideration of the evolution of social ideas and technology in *Erewon* (1872), were undoubtedly also important. 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 plus wildcards, such as GCRs. In Well's 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?*[16]

Wells also wrote two essays about the topic of human extinction, "On Extinction" (1893) and "The Extinction of Man" (1897), though both no doubt draw as much on his literary imagination as his scientific method. Other science fiction authors continued to break new ground on considering the possibilities of a global catastrophe, most notably Arthur C. Clark and Isaac Asimov in the twentieth century and Cixun Liu in the twenty-first century. Indeed, Asimov wrote the first book-length nonfiction treatment, in 1977, of possible ways a global catastrophe could bring about the collapse of civilization or the extinction of humanity.[17]

---

[14] Medwin, T., 1824. Journal of the Conversations of Lord Byron: Noted During a Residence with His Lordship at Pisa, in the Years 1821 and 1822. Henry Colburn.
[15] Wagar, W., 1983. H.G. Wells and the Genesis of Future Studies. World Network of Religious Futurists. www.wnrf.org/cms/hgwells.shtml.
[16] Wells, H.G., 1902. *The Discovery of the Future: A Discourse Delivered to the Royal Institution*. T. Fisher Unwin.
[17] Asimov, I., 1977. *A Choice of Catastrophes*. Hutchinson.

The first wave was, however, constrained by its commitment to story telling and literary success. It focused on global 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."[18] The first wave nonetheless played a crucial role in focusing scientific and public attention on the long-term challenges facing humanity in a hostile universe.

2.2 Second Wave: Concerned Scientists

The second wave of research into GCRs arose from scientists who became concerned about trends and developments in their own fields of study, 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 initial danger identified by scientists was that "radioactive particles" could contaminate the environment, potentially causing a global catastrophe. As Bertrand Russell and Albert Einstein wrote in their 1955 manifesto:

> *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. It is feared that if many H-bombs are used there will be universal death, sudden only for a minority, but for the majority a slow torture of disease and disintegration.*

An important consequence of the Russell-Einstein manifesto was the Pugwash Conferences on Science and World Affairs. This was founded in 1957 by Russell and Joseph Rotblat, a physicist who worked on the Manhattan Project, and it co-won the 1995 Nobel Peace Prize for its "efforts to diminish the part played by nuclear arms in international politics and, in the longer run, to eliminate such arms."[19] Other scientists who worked on the Manhattan Project and who "could not remain aloof to the consequences of their work" established the *Bulletin of the Atomic Scientists* in 1945.[20] Two years later, it created the iconic "Doomsday Clock" to inform "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."[21] Thus, in response to world events, the *Bulletin*'s Science and Security Board moves the minute hand toward or away from midnight, or "doom." The clock was initially set to 7 minutes before 12:00 am, but in 1953 moved to 2 minutes after the United States and Soviet Union detonated the first thermonuclear weapons. The furthest away it has been to midnight was in 1991, following the end of the Cold War.[22] (Today it is set once again to 2 minutes before midnight.)

---

[18] Bostrom, N., 2002. Existential Risk: Analyzing Human Extinction Scenarios and Related Phenomena. *Journal of Transhumanism*, 9.
[19] Nobel Prize Organization. 1995. The Nobel Peace Prize 1995. https://www.nobelprize.org/prizes/peace/1995/press-release/.
[20] *Bulletin of the Atomic Scientists.* 2016. *Background and Mission: 1945-2016*. http://thebulletin.org/background-and-mission-1945-2016.
[21] Benedict, K. 2018. Doomsday Clockwork. *Bulletin of the Atomic Scientists*. https://thebulletin.org/2018/01/doomsday-clockwork/.
[22] *Bulletin of the Atomic Scientists*. 2017. More Bad News for the Nuclear Industry as Westinghouse Files for Bankruptcy. https://thebulletin.org/2017/03/more-bad-news-for-the-nuclear-industry-as-westinghouse-files-for-bankruptcy/.

Worries about environmental catastrophes also emerged after the Second World War, with early studies of the potential environmental catastrophes like William Vogt's *Road to Survival* and Fairfield Osborne's *Our Plundered Planet* (both published in 1948) sounding the alarm about population growth, soil erosion, and environmental pollution. Further concerns about pollution, overpopulation, and resource extraction/contamination in the early 1960s encompassed chemical pesticides, such as DDT, chlordane, and heptachlor. A pivotal study was Rachel Carson's *The Silent Spring* (1962). Carson was a marine biologist 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.*[23]

Several years after Carson's book, Paul and Anne Ehrlich published *The Population Bomb* (1968). It warned about the catastrophic impacts of overpopulation, which the Ehrlich's claimed could lead to "hundreds of millions" of deaths from starvation. In 1972, the Club of Rome, an organization that included scientists, economists, diplomats, government officials, and other influencers from around the world, published a similar report called *The Limits to Growth*. It 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."[24]

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 they could initiate. These would inject soot into the stratosphere that block incoming solar radiation, thus causing global agricultural failures and maybe even human extinction. The result would be what the atmospheric scientist Richard Turco called "the nuclear winter." One of the most prominent scientists who warned about nuclear winter was Carl Sagan. In October, 1983, Sagan published an article in *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.*[25]

This article brought attention to a two-day conference organized by Paul Ehrlich, resulting in *The Cold and the Dark: The World After Nuclear War* (1984). The aim was to examine the "long-term

---

[23] Carson, R., 1962. *Silent Spring*. Houghton Mifflin.
[24] Meadows, D.H., Meadows, D.L., Randers, J. and Behrens, W.W., 1972. *The Limits to Growth*. Club of Rome.
[25] Sagan, C., 1983. Nuclear Winter. *Parade*.
https://www.cooperative-individualism.org/sagan-carl_nuclear-winter-1983.htm.

biological consequences of nuclear war." While Ehrlich was initially skeptical 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*."[26] Shortly afterwards, Sagan published an article on the "policy implications" of nuclear war for *Foreign Affairs* (1983). In it, he 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." The result of Sagan's activism was most likely net positive. For example, the Soviet Premier Mikhail Gorbachev told Ronald Reagan in 1988 that Sagan was "a major influence on ending [nuclear] proliferation."[27]

Research on the nuclear winter phenomenon was spurred in part by a study published in 1980 by Luis and Walter Alvarez. This hypothesized 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 groundbreaking because it threatened the dominant paradigm that had reigned for more than a century: that changes on Earth occur gradually rather than suddenly as the result of global catastrophes. As Trevor Palmer notes, even into the late 1980s, "it was still far from clear whether mass extinctions were real events, rather than artifacts of the fossil record."[28] This changed dramatically with the discovery of the Chicxulub crater on the Yucatan Peninsula in the 1990s, which provided sufficient evidence to convince the scientific community that global catastrophes can occur. During the 1980s, studies of volcanoes suggested that major eruptions could also catapult particles into the stratosphere that block out incoming light. The realization that natural catastrophes can cause mass extinctions in this way was integral to the discovery that anthropogenic factors, like nuclear conflict, could similarly devastate the planet.

By the early 2000s, scientists had identified a wide range of threats to human survival, including threats associated artificial intelligence,[29] nanotechnology,[30] and high energy physics experiments.[31] These were eloquently explored by Martin Rees in his 2003 book *Our Final Hour*. 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 twenty-first century. As later discussed, Rees estimates that the probability of civilization 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 GCR issues, which may prove to be important.

Central themes of the second wave include the possibility of human extinction, the downsides of scientific and technological progress, the need for a world government (or at least for much greater

---

[26] Quoted in Badash, N., 2009. *A Nuclear Winter's Tale: Science and Politics in the 1980s*. The MIT Press.
[27] Francis, M. When Carl Sagan Warned the World About Nuclear Winter. *Smithsonian Magazine*. https://www.smithsonianmag.com/science-nature/when-carl-sagan-warned-world-about-nuclear-winter-180967198/.
[28] Palmer, T., 1999. *Controversy Catastrophism and Evolution: The Ongoing Debate*. Kluwer Academic / Plenum Publishing.
[29] Good, I.J., 1966. Speculations Concerning the First Ultraintelligent Machine. In *Advances in Computers*, 6, pp. 31-88. Elsevier.
[30] Drexler, E.K., 1986. *Engines of Creation: The Coming Era of Nanotechnology*. Anchor Book, New York.
[31] Dar, A., De Rújula, A. and Heinz, U., 1999. Will Relativistic Heavy-Ion Colliders Destroy Our Planet?. *Physics Letters B*, 470(1-4), pp.142-148.

government involvement in the operation of the market and the applications of scientific research), and the moral responsibility of scientists. For example, in a 1948 "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."[32] Others emphasized the role of scientists in informing the public about global risks. The *Bulletin* and 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 (MIT) in 1969 to counteract the "misuse of scientific and technical knowledge presents a major threat to the existence of mankind."[33]

However, the theoretical framework in which such scientists worked was relatively simplistic. It linked discrete exogenous shocks with catastrophic effects; for instance, by considering a simple causal chain from *nuclear conflict* to *rural or urban firestorms* to *stratospheric soot* to *agricultural failures*. We can call this the "etiological approach" to understanding GCRs. Furthermore, concerned scientists have tended to *oppose* measures to reduce our collective vulnerability and exposure to the hazards they believed scientific research might produce and suggest that there is a strong tradeoff, or potential for moral hazard, between such measures and reducing the risks from scientific research. This arose in part from worries that such measures might be ineffective, although it also seems to have reflected a desire that science in general, or at least their research, should only ever be used for beneficial rather than harmful ends. While an admirable position from which to campaign and raise awareness, this presents an unnecessarily limited view for the purposes of risk assessment and risk management.

### 2.3 Third Wave: Techno-Utopians

We date the beginning of the third wave to Nick Bostrom's 2002 paper "Existential Risks: Human Extinction Scenarios and Related Phenomena."[34] This more or less founded a new field of scholarly inquiry and solidified a "step-change" in thinking about catastrophe and the long-term future of humanity. Whereas intellectuals during the first and second waves focused on human extinction and civilizational 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 there are non-extinction collapse scenarios from which humanity could recover—for example, by establishing Earth-independent colonies on Mars or building bunkers underground or in the ocean.[35] On cosmic timescales, a collapse that humanity recovers from may have only a marginal effect on the total amount of value that we create in the future, so that it seems considerably less important from a moral point of view.[36] On the other hand, a future in which civilization continues to flourish, but technological progress plateaus, preventing our descendants from colonizing the universe, might be seen from this point of view as barely better than total human extinction. This 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."[37]

---

[32] Einstein, A., 1948. A Message to the World Congress of Intellectuals. *Bulletin of the Atomic Scientists*, 4(10).
[33] Quoted in: 1969. The Misuse of Science: An Appeal by MIT Scientists. *Bulletin of the Atomic Scientists*, 25(3).
[34] Although other works clearly played an important role, especially Leslie, J., 1996. *The End of the World: The Science and Ethics of Human Extinction*. Routledge.
[35] Turchin, A. and Green, B.P., 2017. Aquatic refuges for surviving a global catastrophe. *Futures*, 89, pp. 26-37.
[36] This point was first made by Derek Parfit, see Parfit, D., 1984. *Reasons and Persons*. OUP Oxford.
[37] Bostrom, N., 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Transhumanism*, 9.

This novel perspective is based on two normative views. The first is *transhumanism*. This contrasts with *bioconservatism* in advocating the use of technology to enhance the human form. The result could be one or more new species of "posthumans" whose core properties of cognition, emotions, or healthspan have been significantly improved. A posthuman could be superintelligent, possess a preternatural ability to control and modulate its emotions, or live indefinitely long lives. In his "Transhumanist Values," Bostrom identifies the central value of transhumanism as "having the opportunity to explore the transhuman and posthuman realms."[38] Although transhumanist themes can be found in the Western tradition dating back at least to the *Epic of Gilgamesh*, it wasn't until the late 1980s and 1990s, facilitated by the Internet, that a community of transhumanists formed.[39] Initially called "extropianism," the philosophy of transhumanism gained a small but influential following among academics based, in particular, at the University of Oxford where Bostrom founded the Future of Humanity Institute (FHI) in 2005.

The second is *utilitarianism*, in particular a "totalist" interpretation. This ethical theory maintains that an act is morally good if and only if it increases the total net well-being in the world. If people have lives worth living, then the larger the population, the greater the well-being. Hence, total utilitarianism implies that humanity should create as much well-being as possible, including through the creation of as many humans with positive well-being as possible. Since far more well-being could be realized in space than on Earth, we should colonize the universe—more specifically, our "future light cone"—as soon as possible, every second of delay results in "astronomical waste."[40] For example, Milan Ćirković calculates that "the number of potentially viable human lifetimes lost per century of postponing of the onset of galactic colonization is" approximately $10^{46}$—or a "1" followed by 46 zeros.[41] Bostrom similarly estimates that, if the Virgo Supercluster contains $10^{13}$ stars and the habitable zone of an average star can sustain $\sim 10^{10}$ biological humans, there could be an incredible $10^{23}$ biological people per century in the Virgo Supercluster. Even more, if our descendants convert whole planets into computers that run high-resolution simulations in which conscious beings live, some $10^{38}$ "sims" could come to exist per century. The result would be an "astronomical" amount of future value, which, from a utilitarian point of view would be very good.

Combining these two views, what ultimately matters isn't avoiding catastrophes *per se*. As Bostrom notes, wars, epidemics, volcanic eruptions, famines, genocides, and so on, are ultimately "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."[42] What matters are scenarios like technological stagnation, irreversible civilizational collapse, and extinction that would keep humanity from creating huge numbers of lives worth living and exploring the posthuman realm. In other words, existential risks are events that prevent us from building a robust techno-utopia. This differs significantly from the second wave's focus on

---

[38] Bostrom, N., 2003. Human Genetic Enhancements: A Transhumanist Perspective. *The Journal of Value Inquiry*, 37(4), pp.493-506.
[39] Bostrom, N., 2005. A History of Transhumanist Thought. *Journal of Evolution and Technology*, 14(1).
[40] Bostrom, N., 2003. Astronomical waste: The opportunity cost of delayed technological development. Utilitas, 15(3), pp.308-314.
[41] Ćirković, M.M., 2002. Cosmological forecast and its practical significance. *Journal of Evolution and Technology*, 12.
[42] Bostrom, N., 2002. Existential Risk: Analyzing Human Extinction Scenarios and Related Phenomena. *Journal of Transhumanism*, 9.

here-and-now suffering caused by nuclear conflict and environmental degradation (although Carl Sagan (1983) did emphasize the importance of thinking about all the future generations that would never exist if humanity were to go extinct). Whereas the second wave was primarily concerned with the loss of human life, the third wave fixated on the loss of human potential. However, this becomes highly problematic because much of the risk facing humanity in the twenty-first century stems from precisely the technologies needed to achieve the transhumanist and utilitarian goals.

Central themes of this wave include transcending human limitations, maximizing value in the long run, building a techno-utopia, and attaining technological maturity. Bostrom (2013) defines the last term as the "attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved." This as the *ultimate aim* toward which humanity should strive, according to Bostrom, since total control over nature and extreme economic productivity would enable the creation of as many posthumans in our light-cone as possible. This is an unambiguously utopian vision of the future, one shared by many others in the third wave. For example, Ray Kurzweil believes that exponential technological development will lead to the "Singularity." This 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." It will be catalyzed 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, thus yielding a magical "world where there is no distinction between the biological and the mechanical, or between physical and virtual reality." He continues:

> *These technological revolutions will allow us to transcend our frail bodies with all their limitations. Illness, as we know it, will be eradicated. Through the use of nanotechnology, we will be able to manufacture almost any physical product upon demand, world hunger and poverty will be solved, and pollution will vanish. Human existence will undergo a quantum leap in evolution. We will be able to live as long as we choose. The coming into being of such a world is, in essence, the Singularity.*[43]

Another theme is *anthropic reasoning*. This concerns how one should reason about one's location in space and time when insufficient data is available. In his 1996 book, John Leslie offers the most detailed defense to date of the "doomsday argument" based on such reasoning. This asks one to reason as if she is 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, say, 100 trillion. Thus, the doomsday argument concludes that we are systematically underestimating the probability of doom.[44] Bostrom later developed these ideas further, arguing in one case that "the doomsday argument is alive and kicking."[45] 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.[46] This has a

[43] Kurzweil, R., 2006. Reinventing Humanity: The Future of Human-Machine Intelligence. https://www.kurzweilai.net/reinventing-humanity-the-future-of-human-machine-intelligence.
[44] Leslie, J., 1996. *The End of the World: The Science and Ethics of Human Extinction*. Routledge.
[45] Bostrom, N., 1999. The Doomsday Argument is Alive and Kicking. *Mind*, 108(431), pp. 539–551.
[46] Bostrom, N., 2003. Are You Living in a Computer Simulation? *Analysis*, 53(211), pp. 243-255.

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).

A primary limiting factor for this strand of research has been its commitment to transhumanism and total utilitarianism. If the aim of existential risk mitigation is to subjugate nature, maximize 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—will likely conclude that existential risk mitigation is absurd, since they do not share these goals, and 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 opportunities to cooperate and engage with wider research communities.

2.4 Fourth Wave: Effective Altruism

The "Effective Altruism" (EA) movement emerged around 2009 with the charity Giving What We Can, founded by Toby Ord. It was inspired by the "global ethics" of Peter Singer, according to which those in wealthy nations have a moral obligation to help people in impoverished countries on the basis that helping someone who lives 10,000 miles away is no less ethically obligatory than helping a child who's drowning in a lake fifteen feet away. The EA movement aims to take this argument a step further by holding that we are obliged not only to help those who are out of sight, but help them in the most *effective* way possible. Thus, the EA slogan: "doing good better."

Initially, the movement focused on researching and then fundraising for effective ways of alleviating global poverty, 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 maximize well-being, and so this cause was joined by the elimination of factory farming (along with other sources of animal suffering) and shaping the far future (to maximize future wellbeing). The reasoning behind the third cause is this: 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 has been called "*longtermism*." In general, effective altruists tended to focus on shaping the future by providing support, both financial and research based, to existing GCR researchers of the third way, although they also helped to expand the community, supporting the establishment of new organizations such as the Centre for the Study of Existential Risk, the Berkley Existential Risk Initiative, and the Global Priorities Institute.

Part of the reason for this is that while not all EAs were transhumanists, the movement's view about value has largely been utilitarian. 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.*[47]

[47] Beckstead, N., Singer, P., and Wage, M., 2013. Preventing Human Suffering. *Effective Altruism Forum*. https://forum.effectivealtruism.org/posts/tXoE6wrEQv7GoDivb/preventing-human-extinction.

EA has also developed its own unique methodology for deciding which projects or causes to allocate resources to, which we will call the "NTI framework." This consists of three questions: (i) How *neglected* is the cause? (ii) How *tractable* is the cause? And (iii) how *important* is the cause? Thus, if a cause is neglected and important but not tractable, it shouldn't be prioritized. Similarly, if a cause is tractable but unimportant, one should focus on other issues. This framework is what led EAs to identify global poverty, factory farming, and the long-term future of humanity as the three best causes. But it can also be used to determine which of the drivers of GCRs researchers ought to focus on, implying that the biggest may not always be the best. This has led many EA longtermists to prioritize solving the "control problem," or the problem of how to build a machine superintelligence whose value system is properly aligned with our "human values." A misaligned superintelligence could, for reasons outlined in Bostrom's *Superintelligence* (2014), spell doom for humanity. 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 international nuclear security and climate change) makes it an area in which resources can be used more effectively. Other EAs, though, have expressed reservations about this focus.[48] Another area in which the EA movement has tended to judge more resources were needed was global catastrophic biological risks, an area that had been paid relatively little attention by previous waves of GCR research. However, EAs have largely downplayed global catastrophic environmental risks (including the risk from climate change) and nuclear risk. For reasons discussed in the final two sections of this paper, we believe that this may have been a mistake based on an over-simplistic reading of the value of "*importance*" (defined below).

Another emphasis among EAs, including longtermists, is "expected value theory." This is the most influential "decision theory" when agents are choosing between actions that lead to outcomes with known probability—that is, decisions "under risk." It states that rational agents should choose the action with the greatest expected value, which is calculated by averaging the probability-weighted desirability of every outcome that an action could produce. It also buttresses arguments for why mitigating existential risk is extremely important. To quote Nick Bostrom, if $10^{54}$ people 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."[49] Put differently, even if there is a 99 percent chance that the proposition "there will be $10^{54}$ future people" is false, it *still* follows that a person who saves the lives of 1 billion humans today is a hero no more than a person who reduces existential risk by a 0.00000000000000000001 percent. Bostrom reiterates this in his 2013 paper on existential risk, arguing that "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."[50]

While findings 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 involving 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

[48] For example, see Matthews, D., 2015. I spent a weekend at Google talking with nerds about charity. I came away … worried. *Vox*. https://www.vox.com/2015/8/10/9124145/effective-altruism-global-ai.
[49] Bostrom, N., 2012. Existential Risk: Threats to Humanity's Future. https://www.existential-risk.org/faq.html.
[50] Bostrom, N., 2013. Existential Risk Prevention as Global Priority. *Global Priority*, 4(1), pp. 15-31.

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.[51]

In 2015, Owen Cotton-Barratt and Toby Ord proposed an alternative conception of existential risk in terms of expected value theory rather than transhumanism and technological maturity. They argued that Bostrom's (2002) definition of existential risk fails to adequately capture catastrophes like a global totalitarian state that oppresses the global citizenry for a period of time but then collapses, thus enabling humanity to continue its quest to maximize value. Hence, they stipulate that "existential risk" should refer to any "event which causes the loss of a large fraction of expected value." The flip-side of an existential *catastrophe*, then, is an existential *eucatastrophe*, or an event that causes a large *gain* in expected value. Examples of such existential eucatastrophes could include designing a value-aligned machine superintelligence or becoming multi-planetary.

The influx of resources and talent into the field of GCR studies brought about by the fourth wave saw it expand dramatically. However, the paradigms of EA still constrained this research in several respects. For example, many people, including most philosophers,[52] reject the *impersonalism* that underlies longtermism. What we should care about, critics say, isn't the potential well-being of currently non-existent (and possibly never-existent) hypothetical possible people, but people who exist right now.[53] As the Oxford 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.*[54]

Relatedly, instead of accepting expected value theory and then concluding that existential risk reduction is very important, one could reinterpret the argument that tiny reductions in existential risk are tantamount to saving huge numbers of current people as a reason to reject this approach. This is not to say that the views held by most effective altruists are wrong, only that they are not nearly 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 persuade others to allocate resources to these causes.

### 2.5 Fifth Wave: Global Systems Thinking

In the past few years, especially since 2018, a fifth wave within GCR research has emerged. Its most salient feature is perhaps a rejection of the "etiological approach" to thinking about existential risks

---

[51] Yudkowsky, Y., 2007. Pascal's Mugging: Tiny Probabilities of Vast Utilities. *Overcoming Bias*, 20th October 2007 (reposted at https://www.lesswrong.com/posts/a5JAiTdytou3Jg749/pascal-s-mugging-tiny-probabilities-of-vast-utilities)

[52] See Bourget, D., and Chalmers, D., 2014. What Do Philosophers Believe? *Philosophical Studies*, 170, pp. 465–500.

[53] See Kuhlemann, K., 2018. Complexity, Creeping Normalcy, and Conceit: Sexy and Unsexy Catastrophic Risks. *Foresight*, 21(1), pp. 35-51.

[54] Srinivasan, A., 2015. Stop the Robot Apocalypse. *London Review of Books*, https://www.lrb.co.uk/the-paper/v37/n18/amia-srinivasan/stop-the-robot-apocalypse?fbclid=IwAR21kIWBtUxua04yJ3C2DzRjagVN9n42r3k7RJU01bv24hxLnri-E7TYZGw.

- that is identifying and assessing risks according to their principal direct cause. According to a seminal paper by Hin-Yan Liu, Kristian Lauta, and Matthijs Maas (2018), the etiological approach is dangerously simplistic.[55] We should think of risks as the product of hazards, vulnerabilities, and exposures. As previously intimated, hazards (also called "threats") are the precipitating cause of a catastrophe, such as a nuclear war or pandemic; vulnerabilities are the inability of critical systems to withstand hazards without incurring damage, such as the loss of food or the collapse of institutions; and exposures are the features of human society that turn this system damage into personal harm, such as our reliance on just-in-time global food distribution networks or institutions of violence suppression.

Whereas most research within previous waves has conflated GCRs specifically with global catastrophic hazards, Liu et al. argue that vulnerabilities and exposures can have globally catastrophic consequences as well. As they observe, "historical studies of civilizational collapses indicate that even small exogenous shocks can destabilize a vulnerable system." To avoid a global catastrophe, we must focus on all three. This is why conflating existential risks with existential hazards, as the etiological approach tends to do, is dangerous: it could lead to a false sense of security. As Liu et al. write, while

> *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 ... of the research agenda "halting" early. In a real-world context, the availability of simple, straightforward "fixes" might even pose a "moral hazard," if policymakers 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.*[56]

Along similar lines, a team of researchers at the Centre for the Study of Existential Risk, led by Shahar Avin, argues that focusing "mainly on tracing a causal pathway from a catastrophic event to global catastrophic loss of life" is inadequate. Instead, researchers should focus on the complex causal interplay between *critical systems*, *global spread mechanisms*, and *prevention or mitigation failures*.[57]

By emphasizing the complexity of existential risk scenarios, both Liu et al. and Avin et al. foreground the importance of global governance for obviating disaster. Governments have significant control over the extent to which critical systems are vulnerable and can limit the extent to which humanity is exposed to hazards that can spread around the world. This requires coordination between governments, and perhaps the establishment of international organizations or treaties that enable states to work together to minimize hazards, vulnerabilities, and exposures by focusing on critical systems, global spread mechanisms, and how efforts to avoid disaster could go wrong. The fifth wave's focus on systems of risk thus gestures back to the second wave, which

---

[55] Liu, H.Y., Lauta, K., and Mass, M., 2018. Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, 102, pp. 6-19.
[56] *Ibid*.
[57] Avin, S., Wintle, B., Weitzdorfer, J., O'hEigeartaigh, S., Sutherland, Wl, Rees, M., 2018. Classifying Global Catastrophic Risks. *Futures*, 102, pp. 20-26.

led to calls for novel forms of governance, including a world government to prevent another world war.

This wave of GCR research is thus significantly different from the previous two waves. It also provides a more nuanced conception of cause prioritization, which abandons 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 actualization: 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 civilizational development. The idea is that a risk A that civilization will almost certainly have to neutralize to survive should take precedence over a risk B that could obtain but might not.[58]

Considering all three properties thus 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 civilization over potentially other drivers of risk, which while neglected and tractable are also more far off, speculative and avoidable. We believe that this nuance is important, but note that it has put some distance between researchers in the fourth and fifth wave of GCR research and meant that some researchers in the fifth wave have struggled to obtain sufficient funding for their work from the EA-based sources that have done so much to grow and support the research community.

## Section 3: The Current State of Research on the Drivers of Global Catastrophic Risk

These five waves of research have produced a non-trivial amount of research into the various drives of global catastrophic risk (GCR). However, in part reflecting the relative timespan over which these waves have taken place, this research has focused on the likelihood of different existential hazards, generally treated as if they were exogenous catastrophic events that might befall humanity. The great body of this work focuses on the quantitative estimation of this risk and it is these studies that we will focus on in this report, because they are the most prevalent and well researched rather than because we believe they provide the most important information. We shall however, in each case, take a critical perspective and consider what these studies can tell us about the drivers of GCRs more fully defined. This section draws on an extensive literature review that was undertaken by one of the authors of this paper together with Tom Rowe of Virginia Tech and James Fox of the University of Oxford, forthcoming in the journal *Futures* as "An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards."

### 3.1 Overall Assessments of the Risk of a Global Catastrophe

Most studies focused on assessing the total level of GCR have relied on two main methods, neither of which is very suitable for risk assessment purposes. These are analytical approaches that derive

[58] Torres, P., 2018. Facing Disaster: The Great Challenges Framework. *Foresight*. 21(1), pp. 4-34.

the level of risk from other facts, such as the "doomsday argument" and informal subjective expert assessments. Examples of the first type include the following studies:

Gott III, J. R. (1993). Implications of the Copernican Principle for Our Future Prospects. *Nature*, 363, pp. 315-319. Predicts that the prior probability that humanity will cease to exist before 5,100 years or thrive beyond 7.8 million years is 5%. This is based on a version of the doomsday argument that uses statistical assumptions about the most likely place of any given observer in the totality of human history to draw conclusions about the most likely length of that history. Note that this study considers the place of an observer in the timespan over which humanity will exist, and not the observer's place within the total human population.

Wells, W. (2009) Human survivability. In: *Apocalypse When?* Springer Praxis Books. Predicts that the annual probability of civilizational collapse was 1%, and the conditional probability of human extinction resulting from this was 30-40%, with this risk decreasing considerably over time. Wells begins with Gott's version of the Doomsday argument, which leads to the conclusion that human history can be expected to last millions of years. However, he argues that Gott was wrong to suggest that the risk of human extinction and civilizational collapse would be evenly distributed across this time period. By considering both the kinds of risks that humanity faces and statistical evidence about the "lifespan" of our economic and cultural institutions, he argues that the risk of civilizational collapse (and human extinction) is presently very high, but that it will have a relatively short half-life, implying that the level of risk will rapidly diminish over time.

Simpson, F. (2016). Apocalypse Now? Reviving the Doomsday Argument. arXiv preprint arXiv:1611.03072. Predicts that "Humanity's prognosis for the coming century is well approximated by a global catastrophic risk of 0.2% per year." This paper offers a more pessimistic version of the doomsday argument. It considers the likely position of a given observer in the population of all past, present and future human beings and revises Gott's arguments in other ways to account of objections to it. The 0.2% per year figure is not taken to be the "conclusion" of this argument, since the risk of human extinction will change over time depending on demographic trends; however, it will remain a good approximation for the next 100 years or so.

Studies such as this are of no value if we conceptualize GCRs in terms of governance failures since they provide no assessment of the drivers of risk, its composition, or how it can be reduced. Indeed, these arguments specifically postulate that the level of risk can be formulated independently of any actions that are taken to manage, or contribute to, such risks. Examples of the second type of study include the following:

Leslie, J. (1996). *The End of the World: The Science and Ethics of Human Extinction.* Routledge. Concludes that "the probability of the human race avoiding extinction for the next five centuries is encouragingly high, perhaps as high as 70 percent." Leslie begins his argument with a version of the doomsday argument, which he sees as implying that human extinction is likely to occur soon. However, he then argues that all of the foreseeable routes that lead to human extinction are unlikely and suggests that our chance of survival is probably higher than this argument on its own suggests.

Bostrom, N. (2002). Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards. *Journal of Transhumanism*, 9. Argues that the probability that "an existential disaster will do us in" is greater than 25%. However, this is a purely subjective judgement based, the author claims, on the balance of the evidence. The methodology for weighing this evidence is not discussed and no time frame for this prediction is given. This is perhaps because, as section 2 discussed, Bostrom is concerned about existential risk in the context of failing to achieve a state of "technological maturity."

Rees, M. J. (2003). *Our Final Century*. Basic Books. Suggests that "The odds are no better than fifty-fifty that our present civilization on Earth will survive to the end of the present century." This is explicitly stated as a subjective best guest. Attention is then drawn to the number of decisions and choices that could impact on human survival, indicating that this assessment relates specifically to the probability that people will make the wrong decisions.

Stern, N., et al. (2006). *Stern Review: The Economics of Climate Change* (Vol. 30, p. 2006). H.M. treasury. Argued that the probability that our world will exist with people who will be affected by our current choices can be assumed, for the purposes of discounting future wellbeing, to decline at a rate of 0.1% per year. Although much quoted, this is presented as no more than a modeling assumption. Stern asserts that 0.1% "seems high" on the basis that "if this were true, and had been true in the past, it would be remarkable that the human race had lasted this long." Yet he also acknowledges that "there is a weak case for still higher levels."

At least one other study builds on this approach by informally surveying a group of experts and aggregating their judgements together:

Sandberg, A. & Bostrom, N. (2008): Global Catastrophic Risks Survey. Future of Humanity Institute, Technical Report #2008-1, pp. 1-5. The median response of an informal survey of 13 participants at the 2008 Oxford Conference on GCR was that there is a 19% chance of overall extinction by 2100. Participants were surveyed on their estimate of human extinction, the death of more than 1 billion people, and the death of more than 1 million people from a list of eight specific threats. However, this list was not taken to be exhaustive.

Another estimate was obtained from an online "prediction market" that allowed users (in this case members of the general public) to place bets on the probability of a global catastrophe:

Metaculus on-line prediction Market - https://www.metaculus.com/questions/578/human-extinction-by-2100/. Asked "will there be zero living humans on planet earth on January 1, 2100?" the median responder gave a 1% credence in this proposition whilst the mean response was 6%. These numbers were taken on the 1st January 2020. These are based on 249 predictions by 207 users in addition to 49 anonymous predictions. The final metaculus prediction will not be released until the question closes on February 9, 2030. Since the market would not be able to pay out until 2100, and nobody would be able to claim their winnings if humanity were to go extinct, participants were all requested to give their responses 'in good faith.'

In theory, if individual opinions like this can be taken as reliable guides to assessing quantitative risk then they can equally well be used regardless of the conceptualization of risk being considered.

However, in practice the majority of these authors treat the overall level of risk as if it could at least be considered as exogenous and reductive concept. The most notable exception is Martin Rees, who in his 2003 book gives considerable attention to questions of governance and governance failure. However, these are presented as part of how we should respond to the existence of risk rather than as the reasons why risk exists in the first place.

Two more informative studies seek to estimate the "background" level of risk, the risk that can most easily be treated as an exogenous feature of human society (although as well will discuss in section 3.6 this is clearly not correct).

Hempsell, CM. (2004). The Investigation of Natural Global Catastrophes. *Journal of the British interplanetary Society*, 57 (1/2), pp. 2-13. Found that "a natural subcritical catastrophe (that would kill over a billion people) would have a 5% to 10% chance of occurring in the next century." This is based on a historical study of events that may have caused the death of more than 10% of the human population. These have a likelihood of causing the collapse of organized society without necessarily causing total human extinction. The historical record for the past 2,500 years suggests that there may have been two such events, the Caldera Volcano in the sixth century, which precipitated a global volcanic winter, and the "Little Ice Age" in the fourteenth century. Both of these events brought about global droughts, famines, epidemics and wars and were probably precipitated by large volcanic eruptions.

Synder-Beattie, A., Ord, T., & Bonsall, M. (2019). An Upper Bound for the Background Rate of Human Extinction. *Scientific Reports*. Data from the archaeological and fossil record about the length of time that humanity has survived so far is employed to estimate an upper bound on the extinction rate from all natural sources combined, including from sources for which we remain unaware. Using only the information that *Homo sapiens* has existed for at least 200,000 years, the probability that humanity goes extinct from natural causes in any given year is almost guaranteed to be less than one in 14,000, and likely to be less than one in 87,000. Using the longer track record of survival for our entire genus *Homo*, the annual probability of natural extinction likely below one in 870,000. The modelling was tested against possible forms of observer selection bias and its conclusions were cross-checked against alternative forms of data, including mammalian extinction rates, the temporal ranges of other hominid species, and the frequency of potential catastrophes and mass extinctions.

However, while these studies are more data driven and relevant than other attempts to quantify general classes of GCRs, they are still clearly hazard driven and do not consider the potential for risk management or the likelihood of its success at preventing such a global catastrophe.

### 3.2 Climate Change

Recent studies of climate change's contribution to GCR have tended to be both vague in their assertions and inconsistent in their assessments of how high this risk is. Many works, such as David Wallace-Well's book *The Uninhabitable Earth* (2019),[59] acknowledge the possibility of a climate induced global catastrophe but shy away from considering it on the basis that doing so involves too much speculation. Others, such as Jem Bendall's widely read working paper, "Deep Adaptation: A

[59] Wallace-Wells, David, 2019. *The Uninhabitable Earth: A Story of the Future*. Allen Lane.

Map for Navigating Climate Tragedy" (2018),[60] go too far in the opposite direction, speculating wildly about disaster scenarios but with no credible consideration of their likelihood or impacts.

To be sure, there are certain widely accepted limits to climate changes survivability. Heat stress itself can pose limits on humanity's ability to survive and this would become a serious issue for certain areas after about 7 degrees of mean global surface warming from pre-industrial levels, and for a majority of the world's population after 11 degrees.[61] Nor are such scenarios impossible. As one risk analysis noted, "on the highest emissions pathway (RCP8.5), a rise of 7°C is a very low probability at the end of this century, but appears to become more likely than not during the course of the 22nd century. A rise of more than 10°C over the next few centuries cannot be ruled out."[62] This would require either that high levels of greenhouse gas (GHG) emissions continue far into the future or that natural feedback mechanisms are stronger than expected. However, heat stress is not the only potentially catastrophic impact from climate change, and other effects, triggered by far smaller temperature increases, could threaten human civilization, either individually or as a group. Several recent studies have attempted to investigate the risk of such impacts.

For instance, one recent study, published in the *Proceedings of the National Academies of Science*, suggests that a global temperature rise of more than 3°C would be "catastrophic" while a rise of more than 5°C would "pose existential threats to a majority of the population" from deadly heat the sea level rise. This conclusion was drawn on the basis that such levels of warming had not been seen within the previous 20 million years. The authors' models suggest that within the next eight decades, there is a 5% chance of exceeding 5 degrees of warming.[63]

Similarly, the 2015 book *Climate Shock* analyzes the uncertainty in standard climate models and finds a 3% chance of passing 6°C under an ambitious "low-medium emissions pathway" and an 11% chance of passing it under a more realistic "medium-high emissions pathway." While stating that we cannot know the full implications of such a temperature rise, the authors describe it as an "indisputable global catastrophe."[64]

A more alarmist, non-peer-reviewed report by the National Centre for Climate Restoration suggested that global temperature increases of more than 3-4 °C "will drive increasingly severe humanitarian crises, forced migration, political instability, and conflict" and "may result in 'outright chaos' and an end to human civilization as we know it," based on a limited

[60] Bendell, Jem, 2018. *Deep Adaptation: A Map for Navigating Climate Tragedy* (IFLAS Occasional Paper 2). Unpublished.
[61] Sherwood, S.C. and Huber, M., 2010. An Adaptability Limit to Climate Change Due to Heat Stress. *Proceedings of the National Academy of Sciences*, 107(21), pp.9552-9555.
[62] King, D., Schrag, D., Dadi, Z., Ye, Q. and Ghosh, A., 2015. Climate Change–A Risk Assessment. Centre for Science and Policy, 8.
[63] Xu, Y. and Ramanathan, V., 2017. Well Below 2 C: Mitigation Strategies for Avoiding Dangerous to Catastrophic Climate Changes. *Proceedings of the National Academy of Sciences*, 114(39), pp.10315-10323.
[64] Wagner, G. and Weitzman, M.L., 2015. *Climate Shock: The Economic Consequences of a Hotter Planet*. Princeton University Press.

scenario analysis. Citing the findings of Reilly et al.,[65] it estimates a 50% chance of crossing this threshold, even if commitments under the Paris Agreement are met.[66]

Other studies offer more optimistic conclusions. For instance, one recent study which assessed climate changes potential to pose an existential threat for the purposes of assessing the costs and benefits of geoengineering research concludes that "the probability of existential catastrophe-level warming is ±3.5%," where this is taken to be warming in excess of 10 degrees Celsius.[67] This is the author's best guess based upon reasonable assumptions about two key factors: (1) the probability of different concentrations of greenhouse gases in the earth's atmosphere, and (2) the conditional probability, given each of these concentrations, that warming may exceed 10 degrees, which is assumed to be the threshold for climate change to pose a direct existential risk to humanity, based on its direct effects. However, the risks associated with climate change's indirect effects and that it could merely cause a global, rather than an existential, catastrophe are presumably somewhat higher, although the author declines to comment on them.

These studies differ in both their assessment of how dangerous climate change is and what the climate thresholds of most concern should be. However, more disconcertingly, they offer little in the way of an answer to the question of why certain levels of climate change could produce a global catastrophe and how we should respond to this. In large part, this reflects the fact that this is a difficult question because answering it requires comprehending complex interacting global systems. Nonetheless, we feel that it also highlights three important limitations of current climate risk analysis:

First, studies tend to assess climate change primarily in terms of the impacts of global mean temperature change, which is only one, albeit important, aspect of this change. Our emission of GHGs is also affecting many other aspects of the global system, including rainfall patterns, ocean acidity, extreme weather, and the balance of energy between the upper and lower atmosphere and the oceans; local effects can differ markedly from the global average. To assess climate change's catastrophic potential, we must think holistically about all of its impacts.[68]

Second, the studies we mention above all build on existing climate and integrated assessment models, albeit not uncritically. These are sophisticated scientific tools that represent our best, albeit still limited, understanding of how climate change will affect human societies. But they are not well suited to studying GCRs. One problem is that existing models, which are best-calibrated for scenarios close to the status quo, are widely acknowledged to perform poorly when applied to more

[65] Reilly, J., Paltsev, S., Monier, E., Chen, H., Sokolov, A., Huang, J., Ejaz, Q., Scott, J., Morris, J. and Schlosser, A., 2015. Energy and Climate Outlook: Perspectives from 2015. MIT Joint Program on the Science and Policy of Global Change.

[66] Dunlop, I. and Spratt, D., 2017. *Disaster Alley: Climate Change Conflict & Risk. Breakthrough National Centre for Climate Restoration*. Melbourne.

[67] Halstead, J., 2018. Stratospheric Aerosol Injection Research and Existential Risk. *Futures*, 102, pp.63-77. The same author keeps an updated, although non-peer-reviewed, open source document of their current assessment of this risk, which they currently see as close to 0, at https://docs.google.com/document/d/1qmHh-cshTCMT8LX0Y5wSQm8FMBhaxhQ8OlOeRLkXIF0/edit#

[68] Briggs, S., Kennel, C.F. and Victor, D.G., 2015. Planetary Vital Signs. *Nature Climate Change*, 5(11), p. 969.

extreme scenarios.[69] More importantly, the greatest risks from climate change are as much, if not more, a feature of its indirect effects on other global systems, rather than the direct effects that climate models are built to study. Current models thus tend to vastly underestimate the potential upper limit for the damage that climate change can cause, and we need new ways of assessing just how great this could be.

Third, as already noted, these assessments move in one direction: from future emissions pathways to climate projections to speculation about what these might mean for humanity and the environment. They ignore how the impacts of climate change will be shaped by human activity, including our reaction to this change. For instance, we might react to rising global temperatures by accelerating the level of our emissions reductions, attempting to geoengineer a safer climate, or merely redoubling our efforts to secure national and individual interests in the face of mounting disasters and dwindling resources. Which of these paths we take will have a significant impact on our species' ability to survive and flourish.

Addressing these three deficiencies requires re-conceiving the assessment of extreme climate change by anchoring our assessments not to the expected level of global temperature rise, but rather to expected social and ecological responses to it. This is true at all levels of climate impact,[70] but takes on even greater relevance when we are considering the issue of GCRs. While linear social and ecological changes can be expected to map on reasonably well to features of global warming like mean temperature rise, social and ecological collapse and other elements of a global catastrophe involving non-linear systemic shifts. Understanding these requires an assessment of their complex causes, which itself demands that we think about the trajectory of climate change in more than just climatic terms. However, such an assessment cannot be carried out without a prior understanding of how and when social and environmental responses might translate into a global catastrophe.

### 3.3 Nuclear War

Similarly, we find that assessments of nuclear war as a driver of GCR provide probably the strongest counterpoint to the above studies since several of these involve the process of "fault tree analysis."

Originally developed to model the emergence of system failures in safety engineering, they have now been widely applied in risk analysis. Fault trees use Boolean algebra (i.e. a logic tree of conditional propositions connected by 'and' and 'or' gates) to model 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 at 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 these probabilities are summed or multiplied 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.

studies that take this approach include the following:

[69] Pindyck, R.S., 2013. Climate Change Policy: What Do the Models Tell Us? *Journal of Economic Literature*, 51(3), pp.860-72.
[70] Travis, W.R., 2010. Going to Extremes: Propositions on the Social Response to Severe Climate Change. *Climatic Change*, 98(1-2).

Hellman, M. (2008). Risk Analysis of Nuclear Deterrence. *The Bent of Tau Beta Pi*, 99(2), 14. Concluded that the annualized probability of a Cuban Missile Type Crisis (CMTC) resulting in World War III is 0.02% - 0.5%. This was given from the following toy model:

$$\lambda_{\text{CMTC}} = \lambda_{\text{IE}} P_1 P_2 P_3$$

$\lambda_{IE}$: Probability of an "initiating event" (a potential first cause of CMTC) is 0.06. There have been three possible initiating events in the last 50 years of nuclear deterrence: Cuban missiles in 1962, Naval blockade of Cuba in the 1980s, and the deployment of American missiles in Eastern Europe. Taking the average rate of these possible initiating events (3 in 50 years), one obtains an annualized probability of an initiating event of 0.06. The category of events seems to be influenced by the author's wish to not be alarmist, and on this basis the paper avoids some possible initiating events, such as the Berlin Crisis of 1961 and the Yom Kippur War of 1973.

$P_1$: The probability of an initiating event resulting in a CMTC. This is set at 0.33 based on the fact that only one of these three initiating events actually *was* the Cuban missile crisis, though the other two clearly had the potential to trigger an event of equal severity had conditions been slightly different.

$P_2$: The conditional probability of a CMTC leading to the use of a nuclear weapon. Since this hasn't happened before, the author relies on the reported *subjective probability estimates* from those involved in the Cuban missile crisis (ranging from 0.01 to 0.5). This is a large range. The author's lower bound of the probability estimate, 0.1, accommodates the fact that the participants stated their estimates before the Russian battlefield nuclear weapons were known in the West. This gives an updated range of 0.1 to 0.5.

$P_3$: The probability that the use of a nuclear weapon results in full scale nuclear war. The author uses reported estimates for this from John F Kennedy and Robert McNamara to arrive at a probability bound of 0.1 to 0.5.

Barrett, A. M., Baum, S. D., and Hostetler, K. (2013). Analyzing and Reducing the Risks of Inadvertent Nuclear War Between the United States and Russia. *Science & Global Security*, 21(2), 106-133. https://doi.org/10.1080/08929882.2013.798984. Concludes, with 90% confidence, that the annual probability of accidental nuclear war between the US and Russia is from 0.001% to 7%. The authors use fault tree analysis to explore possible paths towards the initiation of a nuclear war between the USA and Russia. They then use a range of sources to establish baseline probabilities for each of the fault tree's nodes to produce an estimate for the annual probability of inadvertent war between the United States and Russia (p. 109). It is assumed that the occurrence of mistaken attack indicators are independent random events (p. 113).

Fault trees are especially valuable by providing insights that can be used to study risk mitigation by modeling 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 accurately capture the effect of individual policies or interventions. Seth Baum and Anthony Barrett have since improved their fault

tree analysis with Robert de Neufville, in their working paper "A Model for the Probability of Nuclear War." Baum and Barrett have also extended it to consider the effects of nuclear war in their working paper "A Model for the Impacts of Nuclear War." These two papers offer, we believe, the single best study of a particular driver of GCRs from the governance-focused risk assessment approach that we describe in section 1.

Other assessments of the likelihood of nuclear war, while not employing fault tree analysis, still provide highly useful information. These include:

Lundgren, C. (2013). What Are the Odds? Assessing the Probability of a Nuclear War. *The Nonproliferation Review*, 20(2), 361-374. Finds that "[t]he first sixty-six years of the nuclear age produced a 61 percent chance of a nuclear war." Bayesian statistical reasoning is used to assess the occurrence likelihood of a number of past counterfactual events and the implied probability that a nuclear war could have started. The author notes that this is "an especially applicable mathematical method of calculating probabilities where only limited data are available and assured knowledge is not possible." The value of this approach stems both from the detail in which the author sets out his reasoning and the breadth of evidence they consider, including the opinions of actual decision makers about the likely consequences of their actions.

One important feature to note from all three of the above studies is that they provide a relatively focused risk assessment not of the risk or nuclear war but of either some particular nuclear war scenario (in the case of the first two studies) or of the historical risk of nuclear war (in the third). They also allow for relatively high degrees of uncertainty in their risk assessments, which matches with our own understanding. However, in doing both of these things they provide evidence for a reasonable flaw for the risk of an international nuclear exchange that seems, in our opinion, more than sufficient to warrant taking this risk seriously. While these estimates may look considerably less "sexy" (see section 4) as a result of these qualifications we believe that they are the kinds of findings that should most interest and concern scholars of existential risk.

One key finding from all three studies is the need to focus on humanity's vulnerabilities, both to nuclear war and to the kind of precipitating events that could lead to war. The hazards associated with nuclear war (the mechanics of nuclear weapons themselves and of nuclear winter and other impacts that they might have) and our exposure to them (the damage these impacts would cause to humanity) are relatively well understood. However, a key driver of the level of risk we currently face is determined by features such as the resilience of political de-escalation and conflict prevention institutions and our ability to respond to and recover from a nuclear war should one occur. Given the many difficulties that have been encountered in nuclear disarmament we believe that this points the way not only for important further research and risk assessments, but also the direction that policy interventions aimed at reducing the GCRs from nuclear weapons should take.

Sadly, there remains a lot of research that purports to tell us about the strength of nuclear war as a driver of GCRs yet provide no such insight and are, we believe, significantly less robust and meaningful in their assessments. These include the following:

Project for the Study of the 21st Century. (2015). Experts See Rising Risk of Nuclear War: Survey: https://www.scribd.com/document/289407938/PS21-Great-Power-Conflict-Report. A poll of 50 national security experts from around the world found that there is a "6.8

percent probability of a major nuclear conflict in the next 25 years killing more people than the Second World War." Note that the final prediction was produced by averaging the mean and median responses to this question, which is not a recognized statistical technique.

Good Judgment Project (nonpublic data; referenced by Carl Shulman: http://effective-altruism.com/ea/1rk/current_estimates_for_likelihood_of_xrisk/). Super-forecasters at the Good Judgment Project were commissioned to assess this as part of a set of questions on the future of nuclear weapons by the non-governmental organization OpenPhil. The results were reported on the EA forum as follows: "A median probability of 2% that a state actor would make a nuclear weapon attack killing at least 1 person before January 1, 2021. Conditional on that happening ... an 84% probability of 1-9 weapons detonating, 13% to 10-99, 2% to 100-999, and 1% to 100 or more."

Turchin, A. V. (2008) Structure of the Global Catastrophe: Risks of Human Extinction in the XXI Century. Argues that the risk of extinction due to the consequences of nuclear war, or as a result of a "Doomsday Machine," in the twenty-first century is in the order of 1%. The author estimates the current annual risk of nuclear war to be 0.5% and suggests that should a war take place there is a 50% chance of global civilization degrading to a "post-apocalyptic stage" in which humanity's existence will be vulnerable to other extinction factors. Under such conditions, the probability of human extinction is identified as 10%. However, after 30 years, the risk from nuclear war is assumed to have been made "irrelevant by even more powerful and dangerous technologies." Therefore, Turchin estimates the risk of extinction due to the consequences of nuclear war in the twenty-first century to be 0.75%. A Doomsday Machine is any device, substance, or method that could "destroy all of mankind." The author assumes that the probability of creating and applying a "Doomsday Machine" is 10 times less than the likelihood of conventional nuclear war. However, the conditional probability of extinction if a Doomsday Machine is created is 10 times higher than the conditional probability of extinction if there is a nuclear war. Therefore, the overall risk of extinction due to a Doomsday Machine is the same as from nuclear war.

### 3.4 Pandemics

Another driver of GCRs that has been relatively well-studied is global catastrophic biological risk from human pathogens, both naturally occurring and artificially synthesized. At its best these have applied epidemiological models to study the tail risk distribution of potentially pandemic pathogens. Studies that have taken this approach include the following:

Madhav, N. (2013). Modelling a Modern-Day Spanish Flu Pandemic. *AIR Worldwide*, February, 21, 2013. This made use of the AIR Pandemic Flu Model, which combines demographic, epidemiological, and technological modeling to produce a complete model for pandemic influenza and has been extensively peer reviewed. It found that there is a 0.5-1% annual probability of a "modern day Spanish Flu" event, with similar characteristics to the 1918 pandemic including considerable excess deaths amongst young adults. Such a pandemic would likely cause between 21 and 33 million deaths worldwide. More importantly, however, the model highlights how different trends have tended to have driven this risk both up (including demographic changes) and down (including technological improvements) so as to produce a general neutral effect. The model therefore helps to outline the potential drivers for reducing pandemic risk

Other studies have built simpler toy models that take account of epidemiological and actuarial data but do not seek to provide a complete model of the potential spread and impacts of novel pathogens. These include:

Day, T., André, J. B., and Park, A. (2006). The Evolutionary Emergence of Pandemic Influenza. *Proceedings of the Royal Society B: Biological Sciences*, 273(1604), 2945-2953. Reports that the probability of a pandemic occurring in any given year is 4%. A conservative estimate of the 95% confidence interval for the yearly pandemic probability is 0.7 to 7.6%. This was derived from combining "anecdotal" evidence about the number of influenza pandemics over the past 250 years with more recent data about the expected interval between pandemics emerging. Evidence was combined using a well-defined Bayesian formula set out in an appendix to the paper.

Bagus, G. (2008) Pandemic Risk Modelling. *Chicago Actuarial Association*. Finds that a pandemic of the scale of the Spanish Flu, which caused a ±27% increase in global mortality, occurs around once every 420 years. More severe pandemics, which cause a ±42% increase in global mortality, may have a return rate of 2,700 years. An "actuarial model" is constructed in the form of a severity curve based on historical data for the past 420 years of influenza outbreaks. This was found to approximate an exponential curve which was then extrapolated to estimate the probability and severity of more extreme pandemics. The model takes account of shifting demographic features over time but assumes that pandemics have equal severity across all countries.

Fan, V. Y., Jamison, D. T., and Summers, L. H. (2018). Pandemic Risk: How Large Are the Expected Losses? *Bulletin of the World Health Organization*, 96(2), 129. The annual probability of a severe influenza pandemic (one that increases global mortality by at least 0.1%) is 1.6% and the average impact of such pandemics is a global mortality increase of 0.58% (±40 million fatalities). Severe flu pandemics represent 95% of the costs associated with all pandemic influenza. The historical record was used to estimate the total frequency and severity of all influenza pandemics and to generate likely age-specific death rates as a result of a global pandemic. The U.S. historical age distributions, being the most complete, were used as the template for global age distributions. The authors then model the "expected deaths from pandemic influenza risks" with a highly fat-tailed distribution of mortality (meaning that the vast majority of deaths occur from the most severe pandemics).

While each of these studies considers a different well-defined potential global catastrophe, their findings seem to be broadly consistent with one another. But this should not mislead us into the view that their findings are reinforcing. By treating the occurrence of global pandemics as effectively an exogenous variable the latter three studies give the impression that this is a risk facing humanity more or less regardless of our current political and governance situation. However, the evolution and emergence of novel pathogens is in fact influenced to a significant extent by human activity, from biosecurity efforts to global migration and public health policies. If it was carried out at a different time in the last 50 years is seems likely that the Madhav AIR study would have produced meaningfully different results, allowing it to both track the current strength of biologically driven GCRs but also make meaningful policy prescriptions about how this could be reduced, and why it has not been. The other three studies rely on data that is effectively timeless in that it refers to the long term occurrence of pandemics over many years and cannot take account of

such factors. They would thus yield the same results regardless of what governance policies were currently in operation. Understanding studies within this counterfactual framework immediately exposes their potential value in assessing GCRs as a global governance failure. The fact that the studies may be in rough agreement therefore cannot be taken as evidence for their reinforcing one another's conclusions, because only one of these studies conclusions seeks to reflect the current level of risk and any apparent consistency is therefore a matter of accident rather than representing an underlying regularity in this level of risk.

A topic on which a more governance-focused approach has naturally prevailed is artificially created pathogens. This has been addressed in relation to recent concerns about "gain-of-function" research with the potential to produce novel pathogens of increased pathogenic potential that could escape from laboratories and do consider harm. This risk has been considered by a range of studies:

Klotz, L. C., and Sylvester, E. J. (2014). The Consequences of a Lab Escape of a Potential Pandemic Pathogen. *Frontiers in Public Health*, 2, 116. States that the likelihood of a pandemic, through an undetected lab-acquired infection, "could be as high as 27%" over a 10-year research period. The authors take the annual probability per lab of an escape of a virus through an undetected lab-acquired infection (LAI) to be 2.4%. This statistic is taken from the Department of Homeland Security's risk assessment for a planned National Bio- and Agro-Defence Facility in Manhattan, Kansa. They then assume that a research enterprise will comprise of 10 labs working for 10 years to make a virus. So, across this period, the probability of no escape through a LAI will be 0.088. Therefore, the probability of at least one escape from the enterprise through a LAI will be 91%. This is multiplied by the assumed likelihood, a worst-case scenario, of one LAI leading to a pandemic, 30%, to give the overall prediction.

Lipsitch, M., and Inglesby, T. V. (2014). Moratorium on Research Intended to Create Novel Potential Pandemic Pathogens. *MBio*. Each laboratory-year of Gain of Function research into virulent, transmissible influenza virus might have a 0.01% to 0.1% chance of triggering a global infection via an accidental laboratory escape. Such a pandemic could be expected to kill between 2 million and 1.4 billion people. The risk of a global pandemic resulting from a laboratory escape of influenza is determined from multiplying two different probabilities. The first is the risk of laboratory incidents and accidental infections in biosafety level 3 laboratories in which such research may be conducted (estimated to be between 0.2%, on the basis that four infections have been observed over <2,044 laboratory-years of observation, and 1%, using data from the National Institute of Allergies and Infectious Diseases). The second is the probability that an accidental infection of a working lab could lead to a laboratory escape spreading widely around the world (estimated to be between 5% and 60% according to a range of simulation models, with the authors' own model indicating a 10-20% risk). Noting that "readily transmissible influenza, once widespread, has never before been controlled before it spreads globally," the expected severity of such a pandemic is determined by multiplying the historical infection rate of influenza pandemics (24-38%) by possible values for the case-rate fatality of a novel, virulent influenza strain (1-60%). However, it is unlikely that these two figures vary independently and so simple multiplication is likely to be inappropriate.

Fouchier, R. A. (2015). Studies on Influenza Virus Transmission Between Ferrets: The Public Health Risks Revisited. *MBio*, 6(1), e02560-14. Each laboratory-year of gain-of-function research

into virulent, transmissible influenza virus might have a $2.5x10^{-13}$ to $3x10^{-12}$ chance of triggering a global infection via an accidental laboratory escape. This paper is a direct response to Lipsitch and Inglesby (2014). It argues that their estimates "were based on historical data and did not take into account the numerous risk reduction measures that are now in place in the laboratories where the research is conducted."

The controversy that these competing studies created resulted in a significant rift in bioscience communities and was seen as a significant challenge for assessing the risk that this research posed of leading to a global catastrophe. However, we believe that they also highlight the need to incorporate a governance based framework into risk assessment. While their results are presented as findings about the risks from gain of function research of concern itself, the differences between them are mainly about the differences between approaches to risk management within laboratories and the likely success of the special measures that are undertaken to contain novel pathogens. While the authors do present some differences in their views about the likelihood of a laboratory escape triggering a pandemic and the damage that a pandemic might cause they differ far more in their assessment of laboratory safety. Such safety concerns, however, depend significantly on how the research itself is carried out and how much care researchers take when conducting it. Paradoxically, it seems likely that researchers who believe that the risks from their research are high are likely to be more cautious and take more care, lowering this risk, while the opposite is true. In theory the kinds of laboratory in which gain-of-function research of concern is undertaken are impermeable to pathogen escape. However, we are aware of instances of human failure that circumvent these physical protections. It is thus not possible to separate the issues of risk assessment and risk management in these cases and we feel that there are serious problems with undertaking research (or publicizing its findings) as if we could.

3.5 Artificial Intelligence

In coming decades, there appears to be a high probability of artificial intelligence (AI) surpassing human intelligence. The result could be a global catastrophe, including human extinction. While sometimes seen as the archetypal existential hazard, research on assessing AI as a driver of GCRs remains both sparse and often of quite a low quality. The three best studies of which we are aware are as follows:

Müller, V. C., and Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In *Fundamental Issues of Artificial Intelligence*, pp. 555-572. Springer, Cham. A survey was given to 550 experts with different backgrounds in AI and 170 responded. The four groups that were asked were: (1) participants of the conference on "Philosophy and Theory of AI," Thessaloniki, October 2011; (2) participants of the conferences of "Artificial General Intelligence" in 2012 and (3) "Impacts and Risks of Artificial General Intelligence" in 2012, both of which were held at Oxford; and (4) members of the Greek Association for Artificial Intelligence in 2013. The mean credence of these four groups is that the probability that human level machine intelligence would lead to extinction is 18%.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018). When Will AI Exceed Human Performance? Evidence from AI Experts. *Journal of Artificial Intelligence Research*, 62, 729-754. A survey was conducted of 352 research "experts" (individuals who published at two of the "premier venues for peer-reviewed research in machine learning," 2015 NIPS and ICML conferences). Respondents assigned probabilities to outcomes on a 5 point scale.

Asked whether HLMI would have a positive or negative impact on humanity, the "probability was 10% for a bad outcome and 5% for an outcome described as 'Extremely Bad' (e.g. human extinction)" (p. 4).

Baum, S., Barrett, A., and Yampolskiy, R. V. (2017). Modeling and Interpreting Expert Disagreement about Artificial Superintelligence. *Informatica*, 41(7), 419-428. Reports that the probability of a catastrophe resulting from artificial superintelligence is P ≈ 0.25. Nick Bostrom and Ben Goertzel are both leading thinkers on the topic of artificial superintelligence (ASI), but they disagree about the risk of an ASI catastrophe. A fault tree model individually quantifies the probability of ASI catastrophe resulting from Bostrom's (P ≈ 0.51) and Goertzel's (P = 0.1) arguments. When both experts' arguments were combined, the result was P ≈ 0.25. However, the authors note that "these numbers come with many caveats and should be used mainly for illustration and discussion purposes."

It is worth noting how all three of these studies focused on the reconciliation of present expert opinion about the risk of AGI/ASI, since the nature of this risk makes any assessment of it necessarily speculative. Yet it is often noted that "experts" in this field all face unusually strong incentives either to increase their assessment of risk (because their future funding depends upon this being taken seriously) or to downplay it (because their business model depends, at least in part, on AI development not being significantly obstructed. This can be partially addressed by surveying different groups as both of the large surveys we cite above attempted to do. However, we favour the approach of formally describing, and then seeking to combine, the arguments that different experts give for their assessment. While this is still likely to produce a highly imperfect outcome it has the clear advantage of engaging with experts reasoning and highlighting the conditions that might make optimistic and pessimistic assessments of risk more likely to be correct.

### 3.6 Other Drivers of Global Catastrophic Risk

In the accompanying paper, "An Analysis and Evaluation of Methods Currently Used to Quantify the Likelihood of Existential Hazards," we survey studies concerning several additional drivers of GCRs, namely, asteroid strikes, volcanic supereruptions, space weather, and particle physics experiments (with limited coverage of nanotechnology, ecological collapse and "unknown causes"). There is general agreement that these drivers of GCRs are likely to be considerably less significant than those already described, so we will not list these studies here.

However, one important point needs to be made about many such studies, which is that for those focusing on naturally occurring phenomena such as asteroid strikes and volcanic supereruptions, risk assessments are focused on the occurrence of these phenomena, rather than on the probability of their causing a global catastrophe. There is, to be sure, considerable debate about what size of asteroid or volcanic event would be needed to trigger a global catastrophe but, in general, this is seen as a question for astronomers, volcanologists, and climatologists to answer. However, the risk that such hazards pose to humanity is strongly determined by our level of vulnerability to them, which is in large part a matter of global governance. For instance, David Denkenberger and the Alliance to Feed the Earth in Disasters (ALLFED) are seeking to develop strategies for significantly reducing this vulnerability with regard to food security while organizations like the Electric Infrastructure Security Council (EISC) have been working on similar resilience measures in relation to global electricity supplies. Any assessment of "naturally occurring" drivers of GCRs needs to take account of these measures and include an assessment of our current level of resilience, the potential

to increase this level, and the likelihood that we would do so either prior to, or during, any disaster with the potential to trigger a global catastrophe. In the field of disaster studies, it is now widely recognized that there is no such thing as a "natural disaster"; this insight needs to be incorporated into the field of GCR studies.

**Section 4: Emerging Methods for Assessing Global Catastrophic Risk**

The most recent wave of global catastrophic risk (GCR) research is pioneering new methods and frameworks for risk assessment and management that can help to address some of these deficiencies in the existing research base. We have mentioned some of these in passing already in section 2. However, having now assessed the existing GCR research we will return to consider how some of these methods could be used to overcome this.

4.1 The Classification of Global Catastrophic Risks

The first set of tools we wish to draw attention to are those which have played a definitional role is establishing the fifth wave of GCR research and focus on the classification of GCRs and their different aspects. As previously outlined, two such classification schemes have been published which draw attention to different aspects of GCRs for the purposes of assessment and management.

The first of these classification schemes, is that which was developed by researchers at the Centre for the Study of Existential Risk (CSER). As we previously mentioned, this classifies GCRs 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)."[71]

A *critical system* is one whose ordinary operation plays a crucial role in supporting humanity's ability to survive in its current form. Its safety boundaries are the limit or scale of disturbance that could disrupt its normal functioning and trigger a significant reduction in the support they provide. Avin et al. distinguish between seven levels of critical systems and order these according to their degree of dependence on, and emergence from, one another. These are: (1) physical systems (like the spacetime environment and temperature range); (2) biogeochemical systems (which cycle different elements, like carbon, oxygen, nitrogen, hydrogen, and phosphorus through their various states and locations on our planet.); (3) cellular systems (which support the functioning of all forms of life); (4) anatomical systems (which allow the cells of complex lifeforms to specialize and cooperate effectively); (5) whole organism systems (which determine the behaviour of whole complex organisms, like individual human beings); (6) ecological systems (which govern how organisms interact with one another); and (7) sociotechnological systems (which are the highly ordered institutional, technological and cultural systems that humans have devised specifically to help us in gaining independence from and/or power over other systems). Each contains a set of nested subsystems whose operation is vital to maintaining its overall state. So, for example, within the ecological system there are four critical subsystems: decomposition, food chains, mutualism, and primary production. A failure of any one could result in an ecological disaster.

[71] Avin, S., Wintle, B., Weitzdörfer, J., O'hEigeartaigh, S., Sutherland, Wl, Rees, M., 2018. Classifying Global Catastrophic Risks. *Futures*, 102, pp. 20-26.

When the safety boundaries of critical systems have been breached, thereby causing an abnormal of failed state, this can have cascading effects, potentially spreading disruption both globally and to other systems. The mechanisms of this spread include: "natural global scale" mechanisms, such as changes to the stocks and flows of biochemicals; "anthropogenic network" mechanisms, such as the global energy distribution network; and mechanisms involving biological or informational replication, such as the spread of pests or ideologies. Continuing our previous example: for an ecological disaster to have global rather than merely local effects, there must be a global-scale mechanism capable of spreading the harmful effect to most of humanity.

Classifying risks by their critical systems and spread mechanisms not only provides an analytical tool for studying systemic risks without reducing their inherent complexity, it also helps identify policy levers and other opportunities for mitigation. Prevention or mitigation efforts could keep critical system failures from spreading globally and causing catastrophe. For example, if an organization focused on developing innovative solutions to growing crops without relying on the sun were sufficiently well-funded by wealthy philanthropists, governments, or universities, a supervolcano that blocks out the sun for months or years might not push humanity to the precipice of extinction. Similarly, a lethal pathogen that replicates around the globe would not pose an existential threat if effective vaccines have been developed and distributed to those susceptible to infection. However, GCRs also involve significant challenges to designing and implementing such strategies, and these form the crucial third pillar of this scheme, which identifies four key classes of mitigation fragility: "individual," (such as personal biases) "interpersonal," (such as conflict and reputation management) "institutional," (such as institutional inertia and path dependency) and "beyond-institutional" (such as global cooperation failures and the tragedy of the commons).

The other classificatory schema was developed by Hin-Yan Liu and colleagues (2018) and builds on this framework by further classifying the hazards, vulnerabilities, and exposures that contribute to GCR. They classify vulnerabilities and exposures as ontological, intentional, or related to more complex social arrangements, either passively/indirectly or actively/directly.[72] Ontological vulnerabilities relate to the very nature of human existence (such as the possibility that we exist in a computer simulation and our reliance on a reasonably stable, low entropy, high energy environment to survive); passive vulnerabilities relate to a failure to act (such as a failure to mitigate climate change); active vulnerabilities relate to our optional activities and choices (such as the development of dual-use technologies); and intended vulnerabilities relate to the intention to bring about human extinction (such as implementing a misaligned superintelligence or a mutually assured destruction system of nuclear deterrence). Similarly, exposures can be ontological (such as our exposure to asteroids and supervolcanoes by dint of our existence on planet earth) or intentional (such as our exposure to flood risk and tsunamis by virtue of our desire to live near the sea). Meanwhile, direct exposures arise from particular human activities (such as our exposure to a change in the state of our universe by dint of carrying out certain high energy physics experiments or to terrestrial intelligence by dint of our broadcasting our existence via SETI and other means), while indirect exposures stem from more complex interactions between humanity and the systems that we have come to depend upon (such as the creation of common global infrastructure exposing us to risks from space weather).

Whilst noting that "existential risk = hazard x vulnerability x exposure," so that the etiology of risk

[72] Liu, H.Y., Lauta, K., and Mass, M., 2018. Governing Boring Apocalypses: A new typology of existential vulnerabilities and exposures for existential risk research. *Futures*, 102, pp. 6-19.

will form an essential component of its scientific study, this approach situates the aforementioned "etiological approach" to understanding GCRs within a broader framework of vulnerabilities and exposures, and finds that even relatively small-scale hazards can precipitate an existential catastrophe if the level of vulnerability and exposure is sufficiently high. The only way to achieve genuine safety is not to eliminate hazards, but rather to decrease humanity's vulnerabilities and exposures.

4.2 Connecting Planetary Boundaries and Global Catastrophic Risk

A second suite of conceptual tools has been developed specifically for the assessment of global catastrophic environmental risks. These build on the well-known "planetary boundaries" framework for assessing threats to nine key components of the earth system.[73] This provides a useful tool for studying the broader systemic effects of climate change and humanity's other environmental impacts. Its empirical basis lies in the parameters of relative environmental stability provided by the Holocene epoch in which human civilization has arisen and how we are moving beyond these. While it provides some indication of where environmental threats to human civilization are likely to emerge, it says little about their impact on societies.

For this reason, Seth Baum and Itsuki Handoh (2014) have sought to expand the planetary boundaries concept by creating a framework called "Boundary Risk for Humanity and Nature" (BRIHN). This assesses the risk that crossing planetary boundaries will incur an irreversible loss for humanity. In brief, the BRIHN framework yields a more precise conception of GCR that doesn't, as past concepts have, rely on arbitrary quantifications of human damage. Rather, they understand GCR as the risks "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."[74] The BRIHN approach thus constitutes an early attempt within ERS to redirect the spotlight of scholarly attention away from epistemically neat scenarios and instead analyze how potential disaster scenarios could unfold from a more "systems theory" perspective. The 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). For now, this important framework remains underdeveloped and only informally applied.

A similar, more evaluative approach was taken by researchers at the UCLA Institute for Environment and Sustainability.[75] They argue that crossing a planetary boundary is most dangerous where it is likely to produce reinforcing environmental feedback loops and multiplicative stresses (for instance, when there are multiple spread mechanisms and mitigation fragilities). Yet, they find that most individuals expect global catastrophes to occur through scenarios involving simple and

[73] Steffen, W. et al. 2015. Planetary boundaries: Guiding human development on a changing planet. *Science*, 347(6223), 1259855.
[74] Baum, S., Handohb, I., 2014. Integrating the Planetary Boundaries and Global Catastrophic Risk Paradigms. *Futures*, 107, 13-21.
[75] Kareiva, P. and Carranza, V., 2018. Existential Risk Due to Ecosystem Collapse: Nature Strikes Back. *Futures*, 102, pp.39-50.

direct threats to life. Once again, they found that it is the challenge of responding to emerging risks that underlies their potential to precipitate global catastrophes. They further argue that we already know how to manage complex risks: by supporting heterogeneity, establishing a modular structure, creating redundancy, introducing balancing feedback to counteract reinforcing feedback, and expecting surprises. Yet our lack of understanding of environmental risks could prevent us from deploying these strategies, thereby rendering a global catastrophe more likely.

4.3 Integrating the Assessment of Governance and Global Catastrophic Risk

A final set of tools combines risk assessment with insights from ethics and policy evaluation to assess humanity's likely ability to respond to potential global catastrophes. As we have argued, this is manifestly vital if we are to study GCRs in a way that takes account of their obvious relation to governance failures.

One of the most significant recent contributions to this field was produced by Karin Kuhlemann (2019), who argued that there are a group of "unsexy" risks characterized by (i) declining access to resources that (ii) result from collective action and (iii) can be expected to pass a point at which we will become unable to satisfy the minimum requirements of well-being within the life expectancy of the youngest members of society (what she calls a "threshold of significance"). 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 organized 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 behaviorally 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 behaviors being "supported by attitudinal predispositions to oppose the kind of regulation of individual freedoms that could address the [risks] while curbing free riding."[76]

This includes many risks relating to the ordinary functioning of the human population and our interactions with the environment, including climate change, soil erosion, biodiversity loss, mass unemployment, political fragmentation, and air pollution, all of which she ultimately sees as manifestations of overpopulation.[77] Kuhleman argues that these risks are, by their nature, less compelling than "sexy" risks, which are "neat, quick and techy," such as nuclear war or biotechnological threats (echoing the conclusions of Kareiva and Carranza, 2018), but she adds that focusing exclusively on sexy risks is misguided.

Another development that seeks to open up new avenues of policy analysis relating to GCRs has arisen from a renewed interest in the various kinds of state and nonstate actors who would either

[76] Kuhlemann, K., 2018. Complexity, Creeping Normalcy, and Conceit: Sexy and Unsexy Catastrophic Risks. *Foresight*, 21(1): pp. 35-51.
[77] A similar point is made in Ehrlich, P.R. and Ehrlich, A.H., 2013. Can a Collapse of Global Civilization be Avoided? *Proceedings of the Royal Society B: Biological Sciences*, 280(1754), p. 20122845.

willingly (terror) or accidentally (error) bring about a global catastrophe if only the means were available. In his 1996 book, John Leslie considers a cluster of "risks from philosophy," as he calls them, such as anti-natalism and negative utilitarianism. This attentiveness to ideology was lost with Bostrom's 2002 publication, which fixated—unsurprisingly given transhumanism's fixation on technology—almost exclusively on what we can call *technogenic* rather than *agential* threats. In recent years, though, scholars of GCR have explicitly concentrated on the agent side of the agent-artifact dyad, given that dangerous dual-use technologies almost always (a) require agents or users to cause harm, and (b) are becoming not only more powerful but more accessible to nonstate actors like small groups and even single individuals. This raises the important question of "agential risk," i.e., "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."[78] This can also be understood as our vulnerability and exposure to certain technological hazards that is created by the fact that there are certain groups of people who are much more likely to cause harm with those technologies than others would. There are broadly five basic categories of individuals or groups that give rise to agential risks, including (i) apocalyptic terrorists, (ii) ecoterrorists and violent neo-Luddites, (iii) omnicidal moral actors, and (iv) idiosyncratic actors (including value-misaligned machine superintelligence and suicidal states).[79] However, the question of "what type of individual or group would willingly push an existential-catastrophe-causing 'doomsday button' if only one were within finger's reach?" has become a topic of serious scholarship only since about 2017, meaning that many important insights are yet to be discovered.

While primarily a scholar of the third and fourth waves, Bostrom (2019) has recently proposed a model of GCRs that clearly takes on board some of the methodological insights and concerns we have described here, most notably relating to vulnerabilities and governance failures, in this case related to the "civilizational vulnerability" that arises from our "semi-anarchic default condition."[80] (He defines this as a world order characterized by a limited capacity for preventive policing, a limited capacity for global governance, and diverse motivations among state and nonstate actors.) Under such conditions, Bostrom argues that our civilization is vulnerable to two classes of technological hazard (each of which can be split into two further sub-classes). Although he clearly still views these from a very hazard centric perspective and goes on to label each with an imagined technology that he feels we might be vulnerable too, rather than keeping his definitions focused on the vulnerabilities themselves. The two 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 incentivizes actors to use their powers to cause mass destruction, so that either:
    a. powerful actors can produce civilization-devastating harms and face incentives

[78] Torres, P., 2018. Agential Risks and Information Hazards: An Unavoidable but Dangerous Topic? *Futures*, 95, pp. 86-97.
[79] *Ibid*, Torres, P., 2018. Who would destroy the world? Omnicidal agents and related phenomena. *Aggression and Violent Behavior*, 39, pp. 129-138.
[80] Bostrom, N., 2019. The Vulnerable World Hypothesis. *Global Policy*, 10(4), pp. 455-476.

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 civilizational devastation (worse global warming).

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

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

While we find this model is likely to be useful in understanding the kinds of governance failure that are most likely to lead to a global catastrophe we must note that we do not endorse Bostrom's specific suggestion for what it implies, namely, that mass surveillance should be implemented at a global level to track and control potentially dangerous technological development. Nevertheless, we believe that when viewed alongside the insights of other methods relating to the risks of systemic catastrophes and agential risks, it can help us develop better solutions to this problem.

**Section 5: Recommendations for Future Research and Funding**

As we have shown, GCR research can trace its roots back into the nineteenth century. However, as a coherent discipline it is only a few decades old and is still undergoing rapid growth and transformation. Since the early 2000s, work on the topic has generally focused on the individual, exogenous drivers of GCRs, such as artificial intelligence, biotechnology, and supervolcanoes. However, the last few years have seen a flowering of new approaches, representing what we call the fifth wave of GCR research, which offers a much more comprehensive and sophisticated analysis of GCRs, for the first time identifying GCRs as inextricably linked to governance failures, rather than as mere challenges for governments to overcome. With a deeper understanding of risk classification and analysis, as well as new tools that have enabled progress on assessing global catastrophic environmental risks, the critical importance of (global) governance has more clearly come into focus. This has occurred while the ethical assumptions at the heart of GCR research have also begun to change, and consequently there is a significant risk that funding will not follow these new ideas or allow them to have the full impact they deserve to have. For these reasons, *we believe that the Global Challenges Foundation could play a crucial role in stimulating truly innovative approaches to the study and management of the drivers of GCR by supporting research projects that embrace the emerging research paradigm of the fifth wave*.

But how can organizations like the Global Challenges Foundation identify such projects? One option is to employ a prioritization heuristic that combines the insights of the NTI (neglectedness, tractability, and importance) framework and the Great Challenges framework. We call this the Prioritization Heuristic (pH). On this account, projects focused on governance issues relating to hazards, vulnerabilities, and exposures should be given precedence if the topic of research is relatively neglected and tractable but also significant (in terms of scope), urgent (in terms of the need for action), and ineluctable (in terms of the topic's relation to current trends in our global civilization). To make this more concrete, we provide several candidate research projects that appear to satisfy the pH criteria:

5.1 Global Catastrophic Climate Risk

Although climate change in general is relatively well-understood, its potential contribution to GCRs has been almost entirely neglected as a topic of scientific enquiry (as opposed to social activism). Furthermore, this topic is clearly significant, urgent, and ineluctable given the current trajectory of anthropogenic carbon emissions over the past few decades; whether it is tractable is more difficult to assess. But this is precisely why the "information value" of research on this topic could be very high: at present, we simply do not know either way, and given the immense potential stakes, such knowledge could be extremely useful for scholars to acquire. For example, climatologists have discussed the possibility of "tipping points" and "positive feedback loops" that could, once initiated, result in nonlinear effects. However, these have so far been mostly limited to discussions of tipping points within the climate system itself, rather than relating specifically to the impacts that this will have on humanity. Could there be similar tipping points and feedback loops between the climatic system and other critical systems, such as sociotechnological phenomena like food, health, and global security? Perhaps dwindling resources as a result of climate change—a threat multiplier and threat intensifier—results in civil and interstate wars that both (a) impede carbon emissions reduction efforts, and (b) destroy carbon sinks (like forests), thereby further exacerbating climate change. Or given the immense impact of the 9/11 terrorist attacks that were perpetrated by al-Qaeda (~2.4 trillion dollars spent on two wars and hundreds of thousands of deaths), if climate change increases the probability of global terrorism, as some have argued, then a deeper understanding of this potential causal cascade could be extremely important.[81] We outline a detailed research proposal for this general topic in Appendix C.

As Appendix C shows, we have, together with other researchers at the Centre for the Study of Existential Risk, been working on a research agenda to assess climate change's contribution to global catastrophic risk that draws on many of the methodological insights we describe in section 4. Key research questions within this agenda include:

- What can we learn from past instances of social and ecological collapse and how can we use these to inform our understanding of current levels of resilience? Can we construct a "civilizational boundaries" framework that helps us keep track of how close global civilization is to the probabilistic threshold where it might transition suddenly into a state of collapse?
- What risks do ecological collapse and loss of biosphere integrity pose to human societies, at both the local, regional and global scales? Are we approaching a state of independence form aspects of the earth system that might allow us to engineer our way out of climate induced catastrophes, or is a lack of global resilience and appropriate governance mechanisms for managing global catastrophic environmental risk actually making us more vulnerable to this hazard?
- Can we model feedback loops between ecological and sociotechnological systems using techniques from complex systems models that allow the exploration of many different relationships between different critical systems in these domains? And which kinds of system failure are most likely to cascade into a global catastrophe?

[81] Torres, P., 2016. Apocalypse Soon? How Emerging Technologies, Population Growth, and Global Warming Will Fuel Apocalyptic Terrorism in the Future. *Skeptic*, 21(2), 56-62; Juergensmeyer, M., 2017. Radical Religious Responses to Global Catastrophe. In *Exploring Emerging Global Thresholds: Toward 2030*. Orient BlackSwan.

- Can we use fault tree analysis to map out the various conditions for a climate induced global catastrophe in ways that fully represent all the planetary vital signs that are currently changing due to our emissions of greenhouse gasses, rather than having to measure everything in terms of changes in global mean surface temperature? What is the right balance between a usable model and an accurate representation of systemic complexity for such a fault tree to use for it to be helpful for planning global risk management strategies?
- Which approaches to climate risk management and mitigation does the possibility of climate change significantly contributing to GCR make more or less urgent? At what point might it be worth moving from a strict climate mitigation strategy to a more specific global catastrophe prevention strategy, possibly involving the use of geoengineering techniques, and how can we design "tail-risk treaties" in order to coordinate this kind of decision at the global level in order to ensure that policies are effective and to avoid the risk of "moral hazard" from the appearance that global catastrophes can be avoided leading to a reduction in parties commitment to climate change mitigation?

The Centre for the Study of Existential Risk hopes to complete research into all of these questions within the next few years. However, we are concerned that at present we are one of a very small number of institutes (in some cases the only ones) who are willing to take these research questions seriously as matters for scientific, and not simply political, debate. It is therefore urgent that more organizations (including GCR research organizations like the Global Catastrophic Risk Institute and environmental research organizations like the Stockholm Resilience Centre) are incentivized and supported to enter this research space so that we can contribute in productive discussions and debates and ensure that each other's work is rigorous and replicable. The support of a committed, well resourced, partner like the Global Challenges Foundation is one of the most promising ways in which we can hope to achieve this aim.

5.2 The Study and Management of Agent-Focused ("Agential") Risks

The fact that dual-use emerging technologies like CRISPR/Cas9, cyberweapons, and third-generation laser enrichment techniques (e.g., SILEX) are multiplying the number of state and nonstate actors capable of inflicting global-scale harm makes the topic of "agential risk" significant, urgent, and ineluctable. To date, this topic is highly neglected, although research within fields like "terrorism studies" are to a limited extent relevant. As with global catastrophic climate change risk, it is unclear the extent to which mitigating agential risk is tractable: which governmental policies could reduce the likelihood of malicious agents acquiring dangerous technologies? How is civilization vulnerable and exposed to hazards associated with such agents? Is the formation of a global governing system necessary to combat this threat? Are mass surveillance systems the most effective method of ensuring that agents are unable to unilaterally press a "doomsday button" that destroys civilization? If so, what are the various GCRs associated with global governance and mass surveillance?

Tackling these risks requires both research and reform. We believe that the 2020s are the ideal moment for the United Nations General Assembly to establish a specialized "Omnicide Convention" that proactively codifies into international criminal law the crime of omnicide, i.e., the intentional and unilateral causing of human extinction.[82] This is, we argue, neither a "crime against humanity"

[82] Torres, P., under review. International Criminal Law and the Future of Humanity: Toward a Theory of the Crime of Omnicide. https://docs.wixstatic.com/ugd/d9aaad_fb93025939a543968c72d12c2faa3077.pdf.

(CAH) nor a subtype of "genocide," but a potentially much worse "crime against human potential" (CAHP). The urgency of this project is underlined by the unfortunate history of the 1948 Genocide Convention, which was established after the Nuremberg Trials of 1945-46, thus leading some critics to argue that charging Nazi officials with "genocide" is illegitimate because no such crime had yet been established. This led Raphael Lemkin, who coined the term "genocide" in 1944 and dedicated his life to promoting the idea, to complain as follows:

> in 1933, … the author of the present article [Lemkin] introduced a proposal providing for this type of jurisdiction for acts of persecution amounting to what is now called genocide. Unfortunately, at that time, his proposal was not adopted. Had this principle been adopted at that time by international treaty, we would not now have all the discussions about *ex post facto* law, in relation to crimes committed by the German government against its own citizens prior to this war.[83]

There are thus strong legal and ethical argument for establishing an Omnicide Convention *as soon as possible*, given current trajectories of technological development. Yet efforts to achieve this are virtually non-existent. This top-down, governance-focused approach to mitigating GCRs via the apparatus of international criminal law could thus yield great benefits to humanity if sufficiently supported by funders.

5.3 Establishing Models of Government that Can Outlive the Social Contract

An issue directly related to many contemporary global challenges concerns *algocracy*, or "rule by algorithm." It could be that civilization has become too complex for any human, or group of humans, to govern competently.[84] Alternatively, if mass surveillance is necessary to mitigate agential risk, the risk associated with such programs empowering individual state and non-state actors could surpass the risk associated with malicious agents and emerging technologies. This may point the way to an alternative solution to our semi-anarchic default condition (as described by Nick Bostrom and others above), which could be considerably superior at reducing our vulnerability to GCRs than his proposed regime of mass surveillance. There are strong arguments for such an arrangement that can be built on the following premises:

(1) *The Threat of Universal Unilateralism*: Emerging technologies are enabling a rapidly growing number of nonstate actors to unilaterally inflict unprecedented harm on the global village; this trend of mass empowerment is significantly increasing the probability of an existential catastrophe—and could even constitute a Great Filter.
(2) *The Preemption Principle*: If we wish to obviate an existential catastrophe, then societies will need a way to preemptively avert not just most but all possible attacks with existential consequences, since the consequences of an existential catastrophe are by definition irreversible.
(3) *The Need for a Singleton*: The most effective way to preemptively avert attacks is through some regime of mass surveillance that enables governing bodies to monitor the actions, and perhaps even the brain states, of citizens; ultimately, this will require the formation of a

[83] Lemkin, R., 1946. The Crime of Genocide. *American Scholar*, 15(2), 227.
[84] Bar-Yam, Y., Complexity Rising: From Human Beings to Human Civilization: A Complexity Profile. In *Encyclopedia of Life Support Systems*. EOLSS Publishers; and Torres, P., 2018. Facing Disaster: The Great Challenges Framework. *Foresight*. 21(1), pp. 4-34.

singleton.

(4) *The Threat of State Dissolution*: The trend of (i) will severely undercut the capacity of governing bodies to effectively monitor their citizens, because the capacity of states to provide security depends upon a sufficiently large "power differential" between themselves and their citizens.

(5) *The Limits of Security*: If states are unable to effectively monitor their citizens, they will be unable to neutralize the threat posed by (i), thus resulting in a high probability of an existential catastrophe.[85]

This implies that future global governance may need to be algocratic: given the corruptibility of human beings, we may need to abolish our modern, Westphalian state system and establish a global governing apparatus that is run by a non-corruptible algorithm. This algorithm, via a "post-singularity social contract," would be tasked with monitoring citizens around the world in a "trustworthy" manner, since it would be designed specifically to do so. Yet there are many obvious problems with this idea, and indeed the central thesis of the chapter is not to advocate this solution but to identify a potentially catastrophic vulnerability embedded in current governance systems, given the premises of (a) through (e). Within the pH framework, this topic is both highly neglected and extremely important, and thus funding research on the topic could have a high payoff.

## 5.4 Conclusions

These are three of many possible topics that we believe the pH framework would identify as worthy of greater support. Each presents just the kind of problem that emerging methods and approaches of the fifth wave of GCR research are well-suited to resolving in that they deal with GCR as a complex, systemic, multifaceted issue of governance failures that emerge as much from vulnerabilities and exposures as direct hazards. They thus move towards a more comprehensive conception of GCR management as calling for a sea-change in patterns of global governance, rather than a merely technical problem we can engineer ourselves out of. We believe that addressing these problems, and supporting the development of these methodologies offers the most robust perspective to date on some of the most pressing and consequential challenges facing human civilization this century. We hope that the Global Challenges Foundation will agree.

[85] Torres, P., 2018. Superintelligence and the Future of Governance: On Prioritizing the Control Problem at the End of History. In *Artificial Intelligence Safety and Security*. CRC Press.