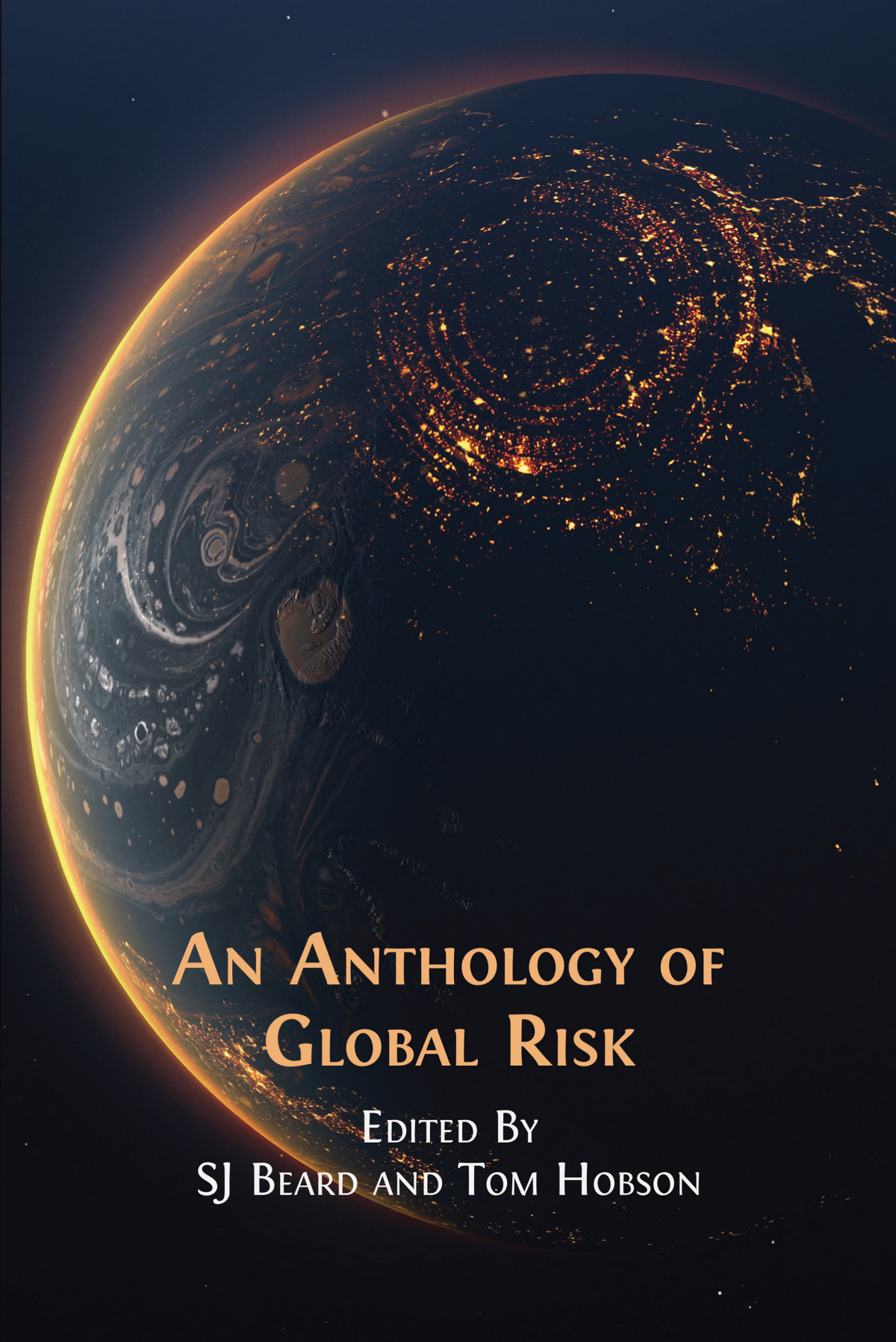

An Anthology of Global Risk

Edited By

SJ Beard and Tom Hobson

https://www.openbookpublishers.com

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

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

ISBN Paperback: 978-1-80511-114-6

ISBN Hardback: 978-1-80511-115-3

ISBN Digital (PDF): 978-1-80511-116-0

ISBN Digital eBook (EPUB): 978-1-80511-117-7

ISBN XML: 978-1-80511-119-1

ISBN HTML: 978-1-80511-120-7

DOI: 10.11647/OBP.0360

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

Cover design: Jeevanjot Kaur Nagpal

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

SJ Beard and Emile P. Torres

Research highlights:

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

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

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

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

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

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

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

key terms, (ii) *motivating values*, (iii) *classificatory systems*, and (iv) *methodologies*.

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

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

Section 1: The Prehistory of ERS

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

1. The scientific community largely rejected the possibility that species could go extinct until the French zoologist Georges Cuvier (1769–1832) demonstrated that elephantine bones unearthed in Siberia and North America belonged to mammoths and mastodons.
2. The belief that an ontological gap separates humans from nature, which was prominent at least until Charles Darwin's *On the Origin of Species*,[5] convinced the scientific community that evolution is a fact about the history of all Earth-originating life, metaphysically integrating humanity into the natural order.
3. Religious eschatologies monopolised thinking about the fate of humanity until the 19th century; it wasn't until the 1960s that the "Age of Atheism" commenced, to borrow a term from Gerhard Ebeling.[6]
4. There was no agreement within the scientific community about the existence of potential kill mechanisms (other than

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

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

Speculative fiction

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

The discovery of the second law of thermodynamics in the early 1850s inspired new thoughts about human extinction among both science fiction writers and working scientists. For example, in his 1870 book *Sketches of Creation*,[10] the American geologist Alexander Winchell

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

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

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

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

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

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

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

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

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

Concerned scientists

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

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

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

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

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

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

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

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

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

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

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

media appearances to increase the potential impact of the research on politicians and the public. For instance, he wrote the cover story for the October 30th 1983 edition of *Parade*, in which he argued that, if a nuclear conflict were to occur:

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

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

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

could occur in the future. During the 1980s, studies of volcanoes also suggested that major eruptions could also catapult particles into the stratosphere that block out incoming light. The realisation that natural catastrophes can induce mass extinctions in this way was integral to the widespread belief that anthropogenic factors, like nuclear conflict, could have similarly devastating effects.

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

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

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

Section 2: Transhumanism, Utilitarianism and the Birth of ERS

While many people's work contributed to the foundation of the field of ERS, most notably John Leslie and Rees;[64] we date the beginning of Existential Risk Studies as a unified field of research to the 2002 paper "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards" by Nick Bostrom (1973–),[65] who trained in philosophy and computer science before establishing the Future of Humanity Institute within the University of Oxford's philosophy department in 2005. This work solidified a number of step-changes in thinking about existential risk and the long-term future of humanity. Whereas previous work had tended to focus on specific catastrophe scenarios or threats, Bostrom's work approached existential risk in a holistic way. Furthermore, whereas previous work focused on human extinction and civilisational collapse, Bostrom focused on catastrophes that would prevent humanity from fulfilling its potential to flourish. Human extinction is the most obvious way this could happen, but it is not the only one. For instance, if human

civilisation collapsed to a state in which we could not recover culturally, economically or technologically this may be almost as bad as if we went extinct completely; even if we were to continue developing but plateau prematurely, before our peak, this could also entail a significant loss of potential for our species. Such considerations led Bostrom to define an existential risk as "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential",[66] which remains perhaps the most canonical definition of the term to date.

Maximising future value

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

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

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

Achieving humanity's potential

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

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

penning a "love letter to humanity", as it were, that time-travels back to the 21st century. As the letter's author puts it, "how can I tell you about Utopia and not leave you mystified? With what words could I convey the wonder?":

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

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

However, this is problematic because much of the risk facing humanity in the 21st century stems from precisely the technologies needed to achieve the goals of transhumanists and singularitarians, making these technologies "dual-use", in that they have the power to both benefit and harm. For example, CRISPR/Cas9 based techniques for gene-editing could potentially halt and even reverse ageing, but could also empower malicious agents to synthesise unnaturally dangerous pathogens. Similarly, hypothetical future devices called

"nanofactories" could usher in an age of unprecedented super-abundance, but could also open the door changing almost any object into any other object at very low cost. Finally, some AI experts have become increasingly concerned that a superintelligent machine could bring about the total annihilation of humanity. As Bostrom, echoing ideas from Yudkowsky, worried in 2002:

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

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

The methodologies of the first wave

This, then, is the intellectual firmament out of which ERS coalesced. If one believes that the future could contain astronomical numbers of super-enhanced posthumans in a galaxy-spanning techno-utopian paradise, then one should care about every possible event that could preclude humanity from achieving that goal. As Bostrom notes, wars,

epidemics, volcanic eruptions, famines, genocides and so on may ultimately be "mere ripples on the surface of the great sea of life" since "they haven't significantly affected the total amount of human suffering or happiness or determined the long-term fate of our species".[80] All that really matters are scenarios like technological stagnation, irreversible civilisational collapse and extinction. This differs significantly from previous concerns, which focused on the process of going extinct and the loss of human life, and, to account for this difference, Bostrom proposes a four-part classification of existential risks according to their outcome. These are:

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

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

The next question for the field of ERS to consider was how this can be achieved. Here, existential risk scholars largely fell back on the methods of their predecessors among concerned scientists (a few of whom, most notably Eric Drexler, became fully part of the ERS community). We referred to this as the "etiological approach"

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

Another methodological feature of this paradigm is the use of *anthropic reasoning* to obtain new information. This concerns how one should reason about one's location in space and time to gain insights into epistemically closed fields of interest, such as predicting the future and understanding other universes. One form of this reasoning is the "doomsday argument", which seeks to assess how long humanity will survive. In *The End of the World*, Leslie offers the most detailed defence to date of this argument.[86] He asks the reader to reason as if they are a random sample of all humans that will ever live. Given that there have existed between 60 and 100 billion people so far (7.8 billion of which are currently alive), the hypothesis that there will be, say, 200 billion in total is much more probable than the hypothesis that there will be 100 trillion, since it is more likely that we are near the middle of human history rather than at one extreme end or the other. Thus, the doomsday argument

concludes that we are systematically underestimating the probability of human extinction in the near future. Bostrom later developed these ideas further, arguing in one case that "the doomsday argument is alive and kicking".[87] Anthropic reasoning also motivated Bostrom's "simulation argument", which purports to narrow down the space of future (and metaphysical) possibility to three scenarios: (i) humanity goes extinct relatively soon, (ii) humanity creates advanced technologies that enable us to run a large number of simulated universes but we choose not to do this, and (iii) we are almost certainly living in a computer simulation.[88] This has a number of real implications for humanity's long-term survival. For example, studies showing that we might not exist in a simulation (or that narrow down the plausible ways that we could be simulated) reduce the probability of (iii), thereby raising the probability of (i), all else being equal. While widely accepted within ERS, these arguments are generally sceptically received by outsiders.

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

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

The second paradigm in ERS built on these foundations, while incorporating insights from the emerging Effective Altruism (EA) movement, which came to be embraced by the vast majority of researchers from the first wave as well as introducing many new people to the field. The EA movement is closely associated with a number of online blogs

such as *Overcoming Bias* (founded in 2006 by Eliezer Yudkowsky and Robin Hanson) and *Marginal Revolution* (founded in 2003 by Tyler Cowen and Alex Tabarrok). It began to take a more substantial form after the Oxford philosopher Toby Ord co-founded *Giving What We Can*, which quickly developed chapters around the world. Ord established *Giving What We Can* after being inspired by the work of Derek Parfit, Peter Singer and others to make a personal decision to give a significant proportion of his income to charities that would most increase well-being, and receiving many enquiries from others interested in doing the same thing.

Doing the most good

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

i. How *Neglected* is the issue?

ii. How *Tractable* is the issue? and

iii. How *Important* is the issue?

Initially, the movement focused on researching and then fundraising for effective ways of alleviating global poverty (as Singer's argument suggested), most notably by fighting tropical diseases, such as malaria. However, as it developed, members raised concerns over whether this really was the most effective way to create value, and so this cause was joined by the elimination of factory farming (along with other sources of animal suffering), shaping the far future (to maximise

future well-being), and most recently tackling mental illness (especially among the poor). The reason many effective altruists decided to focus on shaping the far future is that if one wants to improve the lives of as many people as possible, and if most people who will ever exist will live in the future, then one should focus on the future. This position was most extensively articulated in the philosophy PhD thesis of Nick Beckstead, who called it "longtermism".[90] As Nick Beckstead, Peter Singer and Matt Wage write:

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

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

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

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

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

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

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

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

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

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

In 2015, Owen Cotton-Barratt and Toby Ord proposed a definition of existential risk in terms of Expected Value Theory, which differs markedly from Bostrom's canonical definition from the first wave that was based around the concept of technological maturity. They argued that Bostrom's definition failed to adequately capture catastrophes like a global totalitarian state that oppresses its citizenry for a period of time but then collapses, thus enabling humanity to continue its quest to maximise value. On their view, existential risk should refer to any "event which causes the loss of a large fraction of expected value". This definition also introduces the related concept of an "existential

hope", an event that causes a large gain in expected value; the authors borrow a neologism from J. R. R. Tolkien when referring to the latter events as eucatastrophes.[99] Examples of existential hopes include designing a value-aligned machine superintelligence or becoming multi-planetary.[100]

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

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

quantity of evidence, no matter how great, could ever persuade us to change our mind about it.

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

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

challenges of trying to influence the long-term future of humanity to maximise value.

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

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

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

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

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

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

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

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

Diverging ethical approaches

Alongside efforts to more complex kinds of existential risk, this emerging group of systems thinkers have also pushed back against some canonical normative ideas within previous paradigms. For example, the assumption that developing dangerous dual-use technologies is inevitable as encapsulated by Bostrom's "technological completion conjecture", which states that "if scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained".[114] If the "default outcome" of making a value-misaligned superintelligence is "doom", then why wouldn't humanity be able to put a stop to such research (as we have been able to prevent human extinction through an act of collective suicide or a refusal to procreate)?[115] Another idea that has been scrutinized in recent years

is that space colonisation constitutes an "existential panacea" that will vastly decrease the probability of extinction. For instance, Émile P. Torres (formerly Phil Torres) has argued that there are strong reasons for believing that venturing into space could have catastrophic consequences, likely causing something like an s-risk.[116] Because of these doubts, researchers of the third wave of ERS have tended to pay less attention to what the future of humanity may be like or how to ensure human "flourishing", and have instead focused more on avoiding the present risks facing humanity.

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

Kuhlemann argues that focusing on these risks is wrongheaded. Scholars within ERS need to also consider "unsexy risks" as well. She defines these as dangerous scenarios that could produce an existential outcome, but also have a "high probability of sub-existential outcomes".[118] The three properties of unsexy risks are: first, they are *epistemically messy*, meaning that they "resist precise definition and do not ... map well onto traditional disciplinary boundaries or institutional

loci of governance". Investigating the relevant causal factors and mitigation strategies thus requires "the combination of perspectives from multiple wildly different disciplines, which is a daunting prospect to many researchers and a poor match to how centres of research tend to be organised and funded". Second, they *build up gradually* and hence "play out in slow motion — at least as perceived by humans". This tends to "[obscure] the extent and momentum of accumulated and latent damage to collective goods, while shifting baselines tend to go unnoticed, misleadingly resetting our perception of what is normal". And finally, they are *behaviourally and attitudinally driven* in the sense that their primary causes are "the procreative and livelihood-seeking behaviours constitutive of population growth and economic growth"; these behaviours being "supported by attitudinal predispositions to oppose the kind of regulation of individual freedoms that could address the [risks] while curbing free riding". Examples include phenomena like "topsoil degradation and erosion, biodiversity loss, overfishing, freshwater scarcity, mass un- and under-employment, fiscal unsustainability, and ... overpopulation".

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

The ethical paradigms of this third wave in ERS have also helped inspire more nuanced conceptions of cause prioritisation, which abandon the simplistic notion of *importance* from EA's NTI framework due to its value-ladenness in favour of a more descriptive account of what kinds of challenges most need attention. On one account,

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

Risk classification and the methods of the third wave

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

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

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

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

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

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

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

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

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

1) Technology makes it too easy for individuals or small groups with the appropriate motivation to cause mass destruction, so that it is either:
    a) extremely easy to cause a moderate amount of harm (very easy nukes); or
    b) moderately easy to cause an extreme amount of harm (moderately easy bio-doom).
2) Technology strongly incentivises actors to use their powers to cause mass destruction, so that either:
    a) powerful actors can produce civilisation-devastating harms and face incentives to use that ability (safe first strike); or
    b) a great many actors face incentives to take some slightly damaging action such that the combined effect of those actions is civilisational devastation (worse global warming).

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

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

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

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

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

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

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

2) Response — ensuring that such events do not precipitate global catastrophes, by detecting them early, reducing the time lag between detection and response, ensuring that planned responses won't be stymied by cascading impacts, and identifying leverage points to maximise their impact.

3) Resilience — ensuring that the worst effects of global catastrophes are avoided, by maintaining and increasing the diversity of human settlements and livelihoods, preparing large-scale evacuation and recovery infrastructure, and planning late-stage response measures to deploy under worst-case scenarios.[135]

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

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

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

Section 5: The Future of ERS

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

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

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

may remain unrecognized".[137] If more existential risk scenarios are either actively created ("ontological risk multiplication") or discovered by science ("epistemic risk multiplication"), then the ranks of ERS could further swell.[138]

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

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

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

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

species as a whole, and indeed it is clear that it has created institutions that are as poorly aligned with human values as any superintelligent AI that many of us fear.

Conclusion

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

Acknowledgements

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

Notes and References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

40 Bostrom (2002).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

65 Bostrom (2002).

66 Bostrom (2002).

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

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

69 Sagan (1983).

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

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

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

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

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

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

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

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

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

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

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

80 Bostrom (2002).

81 Bostrom (2002).

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

83 Leslie (1996); Bostrom (2002).

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

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

86 Leslie (1996).

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

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

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

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

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

92 Beckstead (2013).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

118 Kuhlemann (2019).

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

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

121 Bostrom and Ćirković (2008).

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

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

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

125 Baum and Handoh (2014).

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

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

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

129 Leslie (1996).

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

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

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

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

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

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

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

136 Liu, Lauta and Maas (2018).

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

138 Torres (2016); Torres (2017).

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

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

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

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