Transhumanism, Effective Altruism, and Systems Theory: A Brief History of Existential Risk Research

Abstract: This paper offers a detailed examination of the history of existential risk studies. I argue that the field emerged in 2002 with the formalization of "existential risk," and that there have been three distinct paradigms since then. The first was based on utilitarian and transhumanist values, the second grew out of the effective altruism movement, and the third embraces a more systems-theoretic approach to analyzing global risk phenomena. In sketching the evolution of these thought traditions, I offer a critical examination of each, identifying their various strengths and weaknesses. This paper should be of interest to anyone curious about the current state of existential risk research and how the field became what it is today, given its peculiar origins in techno-utopian aspirations and millenarian prognostications.

Section 1. Introduction

According to the *Unique Hazards Hypothesis*, humanity finds itself in the most perilous moment of its 300,000-year history, with perhaps the one exception of the Toba catastrophe.[1] We can define "perilous" as the total amount of *risk potential* that exists at some timeslice, and "risk potential" as, metaphorically, the total amount of TNT strapped to the planet. If this hypothesis is true, it may be unsurprising that a new field focused on the long-term survival of our evolutionary lineage has recently taken shape. This is sometimes referred to as "existential risk studies" (ERS), although some scholars prefer for various reasons "existential risk mitigation" (ERM).[2] For the present purposes, I will use the acronym "ERS" as the appellation of this area of scientific and philosophical inquiry. The aim of this paper is to explore the historical development of ERS and, in doing so, to identify important points of convergence and divergence between different traditions about existential risk. I argue that there are two primary ERS "paradigms"—in a loose sense of this protean Kuhnian term—which I distinguish according to the following issues: (i) *definitions of key terms*, (ii) *motivating values*, (iii) *classificatory systems*, and (iv) *methodology*.

This paper will proceed as follows: section 2 outlines the paradigm that emerged concomitantly with ERS at the turn of the century. It arose from a cluster of techno-eschatologies that anticipated, if not encouraged the active realization of, a utopian world brought about by "technological progress."[3] Section 3 explores the second paradigm—an outgrowth of both the first paradigm and the "effective altruism" (EA) movement, which aims to "combine the heart and the head," as the EA slogan goes. Section 4 examines the rise of "systems thinking" and, with it, an alternative approach to ERS topics that emphasizes the systematicity of global threats, feedback loops, holistic analyses, and scenarios that would culminate in *whimpers* rather than *bangs*.[4] Finally, section 5 concludes with a brief analysis of the value of both contemporary paradigms. This section also briefly considers how thinking about ERS could undergo further metamorphoses in the coming years as the effects of environmental degradation become

---

[1] This occurred some 75,000 years ago, and may have resulted in a severe population bottleneck that came close to causing human extinction.

[2] The reason is that "studies" suggests to some, whether justifiably or not, a field that is not as especially rigorous.

[3] A point worth noting while reading: in a 2009 publication, Nick Bostrom, to some extent at odds with claims he makes elsewhere, writes that "this term ['progress'] has evaluative connotations—of things getting better—and it is far from a *conceptual* truth that expansion of technological capabilities makes things go better. Even if empirically we find that such an association has held in the past (no doubt with many big exceptions), we should not uncritically assume that the association will always continue to hold. It is preferable, therefore, to use a more neutral term, such as 'technological development,' to denote the historical trend of accumulating technological capability."

[4] Note that I do not mean this in the sense of Bostrom 2002.

more salient, new natural risks are identified and anthropogenic risks are created, and ERS becomes less dominated by the limited perspective of white male philosophers.

Section 2: Transhumanism, Emerging Technologies, and Maximizing Utility

Let's begin with two claims. First, the concept of *human extinction*, in the secular-biological sense that is relevant to ERS, is of quite recent provenance, and indeed it would have been virtually unthinkable for most people in the West over long periods of the Common Era. It wasn't until the 1860s, with the formalization of the second law of thermodynamics, that scientists began to take seriously the idea that the human story could one day come to its permanent end in a frozen pond of maximal entropy. Yet widespread worries about near-term human annihilation didn't appear for another century, beginning in the mid-1950s with the Castle Bravo debacle, and a consensus within the scientific community that *nature* poses various existential threats to Earth-originating life only crystallized in the 1980s and 1990s.[5] The point is that the central concept of ERS is quite neoteric.

Second, the historical evolution of the concept of *human extinction* can be periodized into several temporal chunks—*five*, to be exact (author, forthcoming). The most recent step-change in thinking about our collective existential plight in the universe occurred with the founding of ERS by Nick Bostrom's 2002(a) paper, "Existential Risks: Analyzing Human Extinction Scenarios and Related Phenomena." To be sure, this paper drew heavily from a 1996 book by John Leslie titled *The End of the World: The Science and Ethics of Human Extinction*, which offered one of the first comprehensive examinations of not merely *existing risks* associated with phenomena like environmental degradation and nuclear conflict but *emerging risks* arising from molecular nanotechnology and artificial intelligence (AI). Prior to Leslie's book, scholars and organizations like Rachel Carson, Carl Sagan, and the *Bulletin of the Atomic Scientists* tended to select a single threat to raise public and scientific awareness about. But Leslie advanced more than just empirical ruminations about the threats to human existence, he also offered the most rigorous defense up to that point, and perhaps since, of the Doomsday Argument, which concludes that we are systematically underestimating the probability of doom. In doing this, Leslie developed concepts that have become part of the intellectual furniture of ERS, including the *anthropic principle*, *observation selection effect*, and the *self-sampling assumption* (SSA).[6] Bostrom (1999, 2000, 2002b; Bostrom et al. 2010), in particular, has channeled these ideas into ERS, as well as developed the"simulation argument" that purports to narrow down the space of future possibility to three distinct scenarios: (i) humanity goes extinct relatively soon, (ii) humanity creates super-advanced technologies that enable us to run a large number of simulated universes but choose not to do this, and (iii) we are almost certainly living in a computer simulation (see Bostrom 2003). Since (ii) appears unlikely, either extinction is right around the temporal corner or we are almost certainly digital prisoners in a simulated universe.

An especially notable feature of the ERS step-change in thinking about the long-term future was the conceptual innovation of Bostrom's stipulative definition of "existential risk." According to Bostrom, an existential risk is as "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential" (Bostrom 2002). This "lexicographic" definition thus groups together *human extinction* and a class of *survivable* scenarios as the worst-case

[5] See author and co-author, forthcoming.

[6] Note that Brandon Carter, who first articulated the Doomsday Argument, was also known for his early work on the "anthropic principle." Note also that Leslie didn't use the term "self-sampling assumption," nor did he originate the idea, which may have been first discovered by Richard Gott.

outcomes for humanity in the universe—that is, as the outcomes that humanity ought to most avoid—rather than focusing just on the former (see author 2019).[7]

But what motivated this shift to a more inclusive semantic extension? There are two main answers. The first pertains to *ethical considerations* and can be dissected into two distinct sub-theses, namely, the *quantity contention* and the *quality contention*. Taking these in order: the quantity contention concerns the number of future people who could exist within our Hubble volume. This is relevant because from a particular ethical perspective—i.e., the consequentialist theory of utilitarianism—an act is morally right if and only if it increases net well-being, and sentient beings are the "containers" of well-being, to borrow John Rawls's (1974/1975) memorable term. This is what makes utilitarianism *impersonal*. Furthermore, if one adopts an *aggregative function* that rejects both (a) temporal discounting (counting future people less than current people[8]), and (b) a diminishing marginal return on well-being (extra well-being matters less the more well-being there already is), we get the prescription to maximize the total number of people with happy or "worthwhile" lives.

How many people could there be? In a 1983 paper published in *Foreign Affairs*, Carl Sagan calculates that if humanity survives for another 10 million years, there could come to exist some 500 trillion future people. Later, Milan Ćirković (2002) estimated that "the number of potentially viable human lifetimes lost per century of postponing of the onset of galactic colonization is" approximately 10^{46}—or a "1" followed by 46 zeros. Bostrom built upon 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 computronium (or matter that is optimized computer hardware), and if minds are functional types that can be multiply realized by different underlying physical substrates, then we could run vast numbers of simulations in which inscrutable quantities of worthwhile lives come into being. On Bostrom's count, some 10^{38} simulated beings could exist per century in our lonely supercluster—indeed, there are an estimated 10 million additional superclusters in the visible universe.[9] It follows that the future could contain truly astronomical amounts of moral value, given the impersonal and aggregative suppositions above, as well as the (possibly dubious[10]) assumption that such future lives will be worthwhile.

In contrast, the quality contention pertains to the fact that another way of maximizing value is to augment the capacity of sentient beings to experience well-being. That is, one could hold the quantity of persons fixed while increasing the quality of their experience or hold the quality of personal experience fixed while increasing the quantity of experiences (these are not mutually exclusive). The emphasis on qualitative modifications emerged in the late 1980s and 1990s with an intellectual movement called "transhumanism," although the favored term was initially "extropianism" (see More 2003).[11] The central

---

[7] Elsewhere, I offer a comprehensive look at the various definitions that existential risk scholars have employed in the literature; see author 2019.

[8] Temporal discounting makes sense in the case of, say, money, since $100 invested today could yield $110 in a decade, but it doesn't make sense with respect to human lives. As Jason Matheny (2009) points out, following previous scholars, discounting future live yields conclusions that "few of us would accept as being ethical." For instance, discounting "lives at a 5% annual rate, a life today would have greater intrinsic value than a billion lives 400 years hence." Similarly, discounting at a 10 percent annual rate entails that one person today has equal value to an extraordinary 4.96 x 10^{20} people 500 years in the future. That seems patently absurd from a moral perspective.

[9] Although these will fade from view as the universe continues to expand.

[10] See author 2018.

[11] Where *extropy* is meant to contrast with *entropy*. Incidentally, the citation is to a paper written by Max More, who was a prominent extropian. Note that More 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

tenet of transhumanism is that we should use what Mark Walker (2009) dubs "person-engineering technologies" to radically enhance our core biological features, such as cognitive capacity, emotionality, and healthspan (Bostrom 2008). If such enhancements bring about sufficiently radical changes to our phenotypes, the result could be the genesis of one or more species of *posthumans*, on one or more definitions of "species." Indeed, Bostrom writes in his 2003 paper "Transhumanist Values" that the "core value" of transhumanism is "having the opportunity to explore the transhuman and posthuman realms," the reason being that this could hold the key to "realiz[ing] our ideals" in ways that are presently impossible given "our current biological constitution." While the phrase "realize our ideals" may appear rather vanilla, the ultimate goal of transhumanism/extropianism is to usher in a techno-utopian milieu in which everyone becomes superintelligent; minds, whether simulated or not, experience indescribable pleasure; and senescence becomes negligible as a result of either rejuvenation therapies or mind-uploading, thereby rendering people *functionally immortal* (see author, forthcoming).

Consider Bostrom's "Letter from Utopia" (2009/2019), 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 twenty-first century. It describes a paradisiacal posthuman world; 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 pen, I fear, is as unequal to the task as if I had tried to use it against a charging war elephant." The posthuman author continues:

> 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. Jungle and desert and crackling arctic ice; slum and palace and office, and suburban creek, project, sweatshop, and farm and farm and farm, and a factory floor with a whistle, and the empty home with long afternoons. I have sailed on the seas of high culture, and swum, and snorkeled, and dived … 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 buy share [*sic*] one second with you!

Along similarly utopian lines, Ray Kurzweil (2005) anticipates the exponential development of technology bringing about a history-rupturing event known as the "Singularity." Kurzweil is part of a conspicuously optimistic version of transhumanism called "singularitarianism," where a "singularitarian" is someone who "believes that the Singularity is possible, that the Singularity is a good thing, and that we should help make it happen" (Yudkowsky 2000).[12] 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, thus yielding a magical "world where there is no distinction between the biological and the mechanical, or between physical and virtual reality." Kurzweil continues:

> These technological revolutions will allow us to transcend our frail bodies with all their limitations. Illness, as we know it, will be eradicated. Through the use of nanotechnology, we will be able to manufacture almost any physical product upon demand, world hunger and poverty will

---

get better at everything, become smarter, fitter, and healthier. It would be a constant reminder to keep moving forward" (Regis 1994). Note also that many extropians were libertarians; the overlap between libertarians and transhumanists was challenged in 2004 by the democratic socialist James Hughes.

[12] Note that Yudkowsky disavows "everything [I wrote before] 2002 or earlier" (Yudkowsky 2019).

> be solved, and pollution will vanish. Human existence will undergo a quantum leap in evolution. We will be able to live as long as we choose. The coming into being of such a world is, in essence, the Singularity (Kurzweil 2006).

The AI theorist Ben Goertzel (2010) defends a similar normative worldview known as *cosmism*. Focusing more on the cosmos than humanity, it affirms the desirability of pursuing super-powerful advanced technologies that will enable humans to merge with machines, upload our minds, colonize the visible universe, engage in "spacetime engineering," devise new and better ethical systems, and "reduce material scarcity drastically, so that abundances of wealth, growth, and experience will be available to all minds who so desire." The result is that "all these changes will fundamentally improve the subjective and social experience of humans and our creations and successors, leading to states of individual and shared awareness possessing depth, breadth and wonder far beyond that accessible to 'legacy humans'" (Goertzel 2010). At the extreme, future posthumans might even use advanced genetic engineering and nanotechnology to modify the "metabolic pathways of pain and malaise" in organisms to completely eliminate all sentient suffering in the biosphere and beyond, an ideal most comprehensively articulated in *The Hedonist Imperative* (1995) by David Pearce. Harkening back to the arduous struggle to end the enslavement of blacks by whites, Pearce—who incidentally co-founded the World Transhumanist Association (WTA) with Bostrom in 1998—calls this the "abolitionist project."

Returning to Bostrom's definition, the crucial point is that if one believes that the future could contain astronomical numbers of super-enhanced posthumans in a galaxy-spanning techno-utopian paradise, then one should (at least be inclined to) care about *every possible event* that could preclude humanity from reaching "the promised land." The most obvious event that would foreclose the possibility of utopia is human extinction, since it would snap off the *Homo sapiens* twig on the evolutionary branch of Hominini. But there are other possibilities as well. For example, civilization could collapse and never rebuild. It is this range of scenarios that motivates Bostrom's notion of existential risks as events that irreversibly injure our capacity for "desirable future development" (Bostrom 2013). To make this even more precise, one could reconstruct Bostrom's definition in biconditional form as follows: an event X is an existential risk if and only if X would prevent Earth-originating intelligent life from attaining a stable state of *technological maturity* if X were to occur (see author 2019). The technical term "technological maturity" signifies, as Bostrom (2013) puts it, the "attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved." This is significant because it would enable humanity to (a) maximize the total number of people within our Hubble volume, and (b) maximally explore the posthuman realm, thereby satisfying the two desiderata of the quantity and quality contentions.

Bostrom thus proposes a four-part classification of existential risk "failure modes," which I have elsewhere called the "outcome approach" to understanding existential risk (see author 2017). Quoting from Bostrom (2013) and using the term "humanity" in a wide sense to denote our human and posthuman descendants, these are:

> (1) *Human extinction*. Humanity goes extinct prematurely, i.e., before reaching technological maturity.
> (2) *Permanent stagnation*. Humanity survives but never reaches technological maturity.
> (3) *Flawed realization*. Humanity reaches technological maturity but in a way that is dismally and irremediably flawed.
> (4) *Subsequent ruination*. Humanity reaches technological maturity in a way that gives good future prospects, yet subsequent developments cause the permanent ruination of those prospects.

In every case, humanity fails to attain technological maturity in a "stable" manner, or one that would enable us to exploit our cosmic endowment of negentropy *as much as physically possible* and hence to realize the utilitarian and transhumanist objectives specified above. It is this novel emphasis on *potentiality* that constitutes the conceptual innovation of ERS, and which leads Bostrom to formulate a "rule of thumb," the "maxipok rule," to guide any act of impersonal altruism.[13] This states that "the loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole." More formally: "*Maxipok*. Maximise the probability of an 'OK outcome', where an OK outcome is any outcome that avoids existential catastrophe."

But how can one do this? How can one maximize the probability of an OK outcome? This brings us to a central *methodological* feature of the first paradigm, which I have elsewhere called the "etiological approach" to understanding existential risks (author 2017, 2019). It is perhaps the most intuitive conceptualization of risks, its central feature being the individuation of existential risk types according to their primary underlying causes. Example causes include supervolcanoes, asteroids and comets, gamma-ray bursts, solar superflares, natural pandemics, biodiversity loss, mass extinctions, climate change, geoengineering, engineered pandemics, self-replicating nanobots, extraterrestrial invasions, a simulation shutdown, and runaway artificial intelligence, among many others (see author 2017). The reason this approach is methodological is that by mapping out the links from cause to catastrophe, one can devise intervention strategies to wiggle those causes, thereby modulating the effects. One finds this approach in both Leslie (1996) and Bostrom (2002), and it constitutes the organizing principle of Bostrom and Ćirković's 2008 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*. The first subsumes threats like supereruptions and impactors, the second climate change and artificial intelligence, and the third nuclear war and bioterrorism. This etiological approach thus offered ERS a fairly well-defined research program for scholars to pursue: investigate the routes to disaster from trigger, and then root out the triggers to stop the disasters.

The discussion so far has addressed several features of the first paradigm, which coincides with what Karin Kuhlemann (2018) labels the *futurist perspective*. These include: how should "existential risk" be defined? Which values motivate this definition? And how did research methodologically proceed among-first paradigm practitioners? But there is another issue that is integral for understanding why ERS took shape when and where it did rather than at some other moment or place. Simply recognizing that the future could contain huge numbers of people living happy lives, an insight that dates back at least to Henry Sidgwick (1784), may not be sufficient to inspire the formation of a new academic field, one built around the idea of the maxipok rule. Here is the catch: pursuing the utilitarian and transhumanist goals are *themselves* causally linked to the Unique Hazards Hypothesis of section 1. That is to say, filling the universe with sentient beings and exploring the posthuman realm both require the development of extremely powerful technologies, many of which are precisely the technologies responsible for the soaring risk potential of the contemporary world. Why? Because most such technologies are "dual-use" in

[13] This is not to say that there weren't antecedents in the literature that deserve credit. For example, Derek Parfit (1984) 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 civilization manages to rebuild).# And the calculation above from Sagan that 500 trillion future people could exist was propounded in a 1983 article about nuclear winter, which emphasized 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 both focused on extinction in particular, while the original paradigm within ERS recognized that wholly survivable scenarios—even scenarios in which we attain technological maturity—could *still* result in "existentially catastrophic" outcomes, if only from the utilitarian-Sidgwickian "point of view of the universe."

nature. This means, in contemporary parlance, that they can be used for both good or bad ends, to benefit or to harm. For example, the CRISPR/Cas9 system could potentially halt and even reverse aging, but they could also empower malicious agents to synthesize unnaturally lethal pathogens. Similarly, hypothetical future devices called "nanofactories" could usher in an age of unprecedented super-abundance, but they could also swing open the door to "printing out" immensely dangerous weapons at almost zero cost to the user.

Furthermore, some singularitarians in the early 2000s became concerned that a superintelligent machine, if it doesn't solve all of the world's problems, could inadvertently effectuate the total annihilation of humanity. As Bostrom worried in 2002, recapitulating ideas original to Yudkowsky (2001[14]),

> 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.

In sum, the dangers associated with emerging technologies are profound—an apparent fact that Bill Joy eloquently warned about in his famous 2000 article "Why the Future Doesn't Need Us," thus giving them significant public visibility.[15] Yet the only way to hit the utilitarian and transhumanist marks is to develop these very artifacts. This suggests a plausible explanation for why ERS was founded by utilitarians and transhumanists rather than, say, deontologists (who are generally not interested in value maximizing) or bioconservatives (who believe that modifying our "essential" human properties would be wrong). By focusing on the potential benefits of emerging technologies in the late 1990s, the potential harms gradually, and frightfully, came into sight. The discipline of ERS thus emerged from techno-utopian aspirations and millenarian expectations—indeed, skeptics have even ridiculed the Singularity as the "techno-rapture" (see Hughes 2008).

Yet the most dangerous technologies would have been, and will be, developed whether or not transhumanism had coalesced into the intellectual movement it became. So perhaps it is a very good thing that a small group of far-sighted dreamers sounded the alarm when they did rather than much later.

Section 3: Effective Altruism, Longtermism, and Expected Value Theory

The second paradigm in ERS grew primarily from the "effective altruism" (EA) movement, although it also borrowed significantly from the first paradigm.[16] The core tenets of EA took shape circa 2009 especially within the philosophy departments of Princeton University and Oxford University, where Parfit once taught and Bostrom's Future of Humanity Institute (FHI) is anchored.[17] Thus, not only is there

---

[14] Yudkowsky (2001) refers to the risk discussed by Bostrom as "subgoal stomp." As Goertzel (2015) notes in a critical review of Bostrom's *Superintelligence* (2014), "scratch the surface of Bostrom, find Yudkowsky," that is, "nearly all the core ideas of Bostrom's work appeared previously or concurrently in Yudkowsky's thinking."

[15] In the article, Joy, who has sometimes been described as a transhumanist, argues that the hazards posed by advanced technologies—in particular, GNR (genetics, nanotech, and robotics) technologies—are so immense that we ought "to limit development of the technologies that are too dangerous, by limiting our pursuit of certain kinds of knowledge." He suggests 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" (Joy 2000).

[16] Indeed, many researchers within the first now explicitly identify as champions of the second.

[17] And has been since its founding in 2005.

some intellectual proximity between the first and second paradigms, but a degree of physical proximity as well.

The main axiological commitment of EA is to maximize one's positive impact on the world given one's limited access to valued resources. It was inspired in part by the "global ethics" of Singer—a staunch utilitarian based at Princeton—according to which helping a sentient being that lives 10,000 miles away is no less ethically obligatory than helping a child fifteen feet away who's drowning in a lake. If one accepts this intuition, the obvious next question to ask is: how can I best help those who are in need and out of sight, but ought not be out of mind? The EA answer is to consider three factors, a framework first developed by the Open Philanthropy Project: (i) how *neglected* is the issue? (ii) How *tractable* is the issue? And (iii) how *important* is the issue? These criteria enable one to identify which issues are worthy altruistic causes, the list perhaps including making cars safer, improving the institution of science, developing cognitive enhancements, fighting human trafficking, and reforming the US criminal justice system. However, the primary initially identified as most important by charity evaluator Giving What We Can (GWWC), founded in 2009, was *fighting extreme poverty*. This was later joined by two additional causes: *reducing animal suffering* and *improving the long-term future* of humanity, as specified by GWWC's (now) parent organization, the Centre for Effective Altruism (CEA), which shares office space with FHI.[18] Hence, this way of prioritizing causes has led many EAs to conclude that one of the most effective ways of being an altruist is to engage in activities that could increase the probability of "things going well" in the long run, a position now called "*longtermism*." Let's break down this line of reasoning in more detail.

First, there is hardly a debate that longtermism is neglected. As mentioned in section 2, the very idea of *human extinction* is a quite recent addition to our shared library of concepts. Without delving into excessive detail, the reader may find the four primary reasons for this, developed in detail elsewhere (see author, forthcoming) of interest:

> (a) Religious eschatologies monopolized thinking about the future of humanity until the nineteenth century. It was this century that the seeds of secularism, planted by eighteenth-century Enlightenment *philosophes* like Denis Diderot, began to sprout, at least within the intelligentsia. Yet it wasn't until the 1960s that, to borrow a term from Gerhard Ebeling, the "age of atheism" commenced (Hyman 2010), a point germane to the fourth reason below.[19]
> (b) The scientific community almost unanimously rejected the *possibility* that species could go extinct until 1800 and shortly after. This is when the French zoologist Georges Cuvier published an article in which he demonstrated beyond a reasonable doubt that elephantine bones unearthed in Siberia and North America belong to species no longer existing on Earth—mammoths and mastodons. Quite suddenly, expert opinions changed and extinction was seen as biological reality.
> (c) The belief that an ontological gap separates humans from all other animals was prominent at least until Charles Darwin's *On the Origin of Species* (1859). This comprehensive tome convinced the scientific community that evolution is a fact about the history of Earth-originating life, and in doing so *metaphysically* integrated humanity into the natural order.[20] This was an

[18] Note that effective altruism did not emerge *directly* from prior thinking about existential risks, although it did have some early links to transhumanism. For example, Ord co-authored an article with Bostrom in 2005 about the transhumanist desideratum of cognitive enhancement. See Bostrom and Ord 2005.
[19] Thus, there is a significant sense in which ERS emerged not just from the intellectual traditions of utilitarianism and transhumanism, but from atheism as well.
[20] That is to say, Darwin's theory of evolution by natural selection implied that there does not exist an ontological gap between *Homo sapiens*, as Carl Linneaus named us in his *Systema Naturae*, and the rest of the Animal Kingdom. Hence, if *animals* lack an immortal soul, then so do we.

> important development because if one believes that biological species can go extinct but not that *Homo sapiens* is a biological species, then one will have trouble accepting the possibility of human extinction.
> (d) Finally, there was no agreement within the scientific community about the existence of "kill mechanisms" (other than entropy) that could have plausibly annihilated humanity until the second half of the twentieth century. Following the Castle Bravo test in the Marshall Islands, fears of radioactive contamination around the world triggered panic that genetic mutations could cause the human race to degenerate. However, as intimated in section 2, it wasn't until the 1980s and 1990s that scientists devised the nuclear winter hypothesis and finally accepted that natural phenomena could cause mass extinctions.[21] In a phrase, it wasn't that long ago that one would have been scientifically justified in believing that we live on a very safe planet in a very safe universe. We now know this to be false: cosmic and terrestrial assassins abound.

The point of these details is to underline the fact that very little intellectual energy has focused on probing our existential predicament. This is true even in the past three decades, as Bostrom (2013) affirms by observing that far more scholarly papers have been published about dung beetles than human extinction. Similarly, I have previously noted that

> as of January 24, 2018, there were exactly 1,910 results for the word "existential risk." In comparison, there were 2,060 results for "Super Mario Brothers," 2,100 for "dog flea," 2,320 for "French cheese," 8,760 for "anal penetration," 12,800 for "FOXP2," 66,800 for "bicuspid," and 170,000 for "hospitality management"—all of which are dwarfed by the 5,390,000 results for "cancer" (author 2018).

Second, there are at least some reasons for thinking that improving the long-term future is nontrivially tractable. The most obvious way to affect the far future of humanity—to affect intelligent life, if it still exists, literally billions of years from now—is to reduce the probability of extinction. Strategies to do this are readily adducible by examining the causes specified by the etiological approach. Examples include mitigating climate change, halting biodiversity loss, campaigning against nuclear proliferation, solving the AI control problem, and so on. While this had been the default methodology of the first paradigm, the EA-driven second paradigm expanded the focus from what Nick Beckstead (2013) called "targeted" strategies to more indirect interventions for altering the developmental trajectory of civilization; Beckstead calls these "broad" strategies. These could be quite quotidian acts like "improving education, improving parenting, improving science, improving our political system, spreading humanitarian values, or otherwise improving our collective wisdom as stewards of the future" (Beckstead 2013). The focus on a wider range of actions further buttressed the view that improving the far future could be tractable.

Third, one could argue that preventing humanity from going extinct or civilization from collapsing is extremely important from the perspective of many different value systems. For example, every mainstream ethical theory, namely, virtue ethics, contractarianism, deontology, and consequentialism, maintains that *causing* (and maybe also *allowing*) human extinction to occur would constitute a profound moral wrong (see author and co-author, forthcoming). Hence, we have a moral duty to defend against omnicide, or the intentional killing of all humans, which I have elsewhere argued should receive it own specialized convention within international criminal law (author, forthcoming).[22] But not

---

[21] Independently, Tom Moynihan identifies virtually the exact same factors in his compelling and erudite dissertation, "Existential Risk and Human Extinction: An Intellectual History" (2019).
[22] See author, forthcoming, for a detailed proposal to integrate omnicide into international criminal law via an Omnicide Convention.

only are there be reasons to hold that *going extinct* would be very bad, some value-theoretic positions claim that *being extinct* would itself be catastrophic. The strongest case, if sound, for this comes from utilitarianism. On this view, the badness of human extinction *skyrockets* when humanity becomes functionlly extinct, where "functional extinction" refers to the point at which a species is no longer capable of carrying on, perhaps because of insufficient genetic diversity. Although one need not be a utilitarian to be an EA, many EAs are utilitarians or at least are "most sympathetic to utilitarianism," as Will MacAskill asserts (DS 2019). Indeed, the founder of GWWC Toby Ord argues that utilitarianism, along with the Scientific Revolution and Enlightenment, has "greatly contributed to the upbringing of effective altruism" (Ord and MacAskill 2016). And before the fledgling EA movement voted on the appellation "effective altruism," it seriously considered the appellation "effective utilitarian community" (EUC) (MacAskill 2014).

Whatever the exact prevalence of utilitarianism within EA, the basic idea finds expression in the *long-term value thesis* (LTVT), which undergirds longtermism. This is premised on the same calculations discussed in section 2: Earth will remain habitable for another ~1 billion years, there are 100 billion planets in the Milky Way galaxy that could be colonizable, social and technological progress "will let people have much better and longer lives in the future," the protons that comprise matter won't decay for another 10^{40} years, and so on (Todd 2017; Adams 2008). Here the focus is broader than sentient life; it concerns maximizing *whatever one values* in the world, be it art, music, poetry, science, sports, romance, and so on. 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" (Todd 2017). Yet, as Benjamin Todd argues, 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. So, even if you're not sure you directly should act on the thesis, it might still be the most important area for research (Todd 2017).

The EA research community has also been greatly influenced by *expected value theory* (EVT) and *Bayesian probability*. This is partly the result of its intermingling with the so-called "rationalist community," which accrued around the website *LessWrong* that (the reformed singularitarian) Yudkowsky started in 2009. As a colleague of mine once put it, EA is the logical outcome of (a) rationalism's emphasis on rationality, decision and game theory, formal modeling, avoiding cognitive biases, and other epistemological topics, plus (b) Singer's global ethics insight that it matters not how near or far someone is to be morally obligated to help them. For reasons of limited space, I won't explain either EVT or Bayesianism. Suffice it to say that EVT has been employed in arguments for why shaping the far future is the most important thing one could do. For example, Bostrom—an exponent of EA, and hence now a scholar in the second paradigm—argues that if 10^{54} people could come to exist in the future, then "a mere 1% chance of [this estimate] being correct" implies that "the expected value of reducing existential risk by a mere *one billionth of one billionth of one percentage point* is worth a hundred billion times as much as a billion human lives" (Bostrom 2012). In other words, even if there is a 99 percent chance that the prognostication "There will be 10^{54} future people" is incorrect, it still follows that someone who saves the lives of 1 billion humans today is no more a hero than one who mitigates existential risk by a negligible 0.0000000000000000001 percent. Or as Bostrom puts it elsewhere, "even the tiniest reduction of existential risk has an expected value greater than that of the definitive provision of any 'ordinary' good, such as the direct benefit of saving 1 billion lives" (Bostrom 2013). Existential risk reduction thus *overwhelms* the decision calculus because the stakes are so high: future value that could be *astronomically* large.

The centrality of EVT to longtermist theorizing is perhaps most conspicuous in an alternative definition of "existential risk" proposed by Owen Cotton-Barratt and Ord (2015). On this account, existential risks aren't events that can only happen once in a species career, as is the case with Bostrom's definition, but multiple times. This is because Cotton-Barratt and Ord stipulate that an existential catastrophe is any "event which causes the loss of a large fraction of expected value." I have elsewhere argued that this has several practical, if not theoretical, advantages over Bostrom's definition (author 2019). It also implies that just as there are events that could hugely *reduce* expected value, there could also be events that massively *increase* it. The authors borrow a neologism from J.R.R. Tolkien when referring to the latter events as "existential eucatastrophes." Thus, humanity should strive not only to avoid catastrophes but to actualize eucatastrophes, examples of which might be designing a value-aligned machine superintelligence, becoming multi-planetary, or more generally passing through any "great filter" that blockades the road toward maximizing value to the physical limits.[23]

This more complex focus on both achieving great future goods and preventing great future bads marked a shift away from focusing exclusively on building utopia. But the pendulum swung even further for some EAs, most notably those affiliated with the Foundational Research Institute (FRI). These scholars champion what they call a "suffering-focused ethics," which is closely related to the ethical theory of *negative utilitarianism* (see Tomasik 2015/2018; Gloor 2016).[24] On their view, what longtermists should worry about isn't maximizing net good but minimizing total suffering. This led David Althaus and Lukas Gloor (2016/2019) to proposed the idea of a "suffering risk," or "s-risk," which refers to any scenario in which suffering exists "on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far."

For example, imagine a future in which our progeny become posthuman, colonize the universe, and attain a stable state of technological maturity, thus creating vast amounts of well-being. Does this constitute an existential win? On Bostrom's view it surely does. But there is an important datum missing: how much total suffering exists in this world? If the answer is "an astronomical amount," then our progeny will have realized a suffering risk—call it an "s-catastrophe." Given the exclusivist focus on suffering, this group of EAs wants to prioritize the mitigation of s-risks rather than pursuing utopia, perhaps in accordance with what I have called the "minipnok rule," which inverts the maxipok rule by asserting that whenever one acts out of an impersonal concern for humankind as a whole, one should minimize the probability of a "not-OK outcome," where a not-OK outcome is any outcome that fails to avoid an s-catastrophe. 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 among both secular and religious people about what would count as dystopia: pervasive and intense pain, misery, dejection, anguish, unfulfilled desires, ignorance, loneliness, insecurity, violence, oppression, war, genocide—in biblical phraseology, a "weeping and gnashing of teeth." Hence, bad outcomes provide a far less ambiguous target to strike than good outcomes, since agreement converges more on the former. One could therefore argue that we should focus on avoiding hell rather than, as the transhumanists do, reaching heaven.[25]

As of this writing, a large portion—maybe a significant majority—of ERS scholars are EAs with longtermist convictions. Some participated in the first paradigm but now identify more with EA and longtermism than transhumanism, at least publicly, although utilitarianism remains a prominent component of EA theorizing. Not only has this second paradigm introduced its own methodological

---

[23] See Hanson 1998.

[24] A view championed by the WTA co-founder, Pearce.

[25] 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."

toolkit, but it has foregrounded decision-theoretic concepts like *expected value*. The EA movement has also attracted a sizable fanbase among wealthy and super-wealthy donors, including Sam Harris, Peter Thiel, and Elon Musk, all of whom are concerned about the long-term future of humanity (if only because some hope to upload their minds, live forever, acquire superhuman intelligence, traverse the galaxy, and live on exoplanetary colonies[26]). The influx of money has significantly contributed to the movement's cultural notability, since money confers status and status confers popularity; more to the point, a hefty chunk of change has consequently ended up in the pockets of existential risk researchers, which is one reason for CEA and FHI's close affiliation. Unfortunately, this has also meant that criticisms of EA and ERS have often been self-censored: criticizing the hand that feeds is tantamount to biting it. Nonetheless, the second paradigm offered much that the first paradigm lacked, although it remains to be seen whether longtermism has intellectual staying power.

Section 4: Complexity, Systems, Boundaries, and Pluralism

Perhaps the most prominent feature of the third paradigm is the diversity of viewpoints on issues like how to classify existential risks, what the best methods for studying them are, which values should motivate risk analysis, and whether previous definitions of "existential risk" are conceptually and empirically adequate. Underlying this mosaic of opinion is a general emphasis on the *complex systematicity* of existential risks, that is, the jumbled tangle of interacting systems, causal cascades, tipping points, feedback loops, synergistic effects, and slow-moving scenarios that can creep up from behind, all of which may contribute to a catastrophically bad outcome for humanity.

An excellent starting point for understanding this paradigm is Kuhlemann's (2018) discussion of "sexy" and "unsexy" risks. She begins by observing 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*, meaning that, for example, it is not difficult to identify which academic fields 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 *technologically driven* and, as such, "have a close relationship with rather flattering ideas about human ingenuity and intellectual prowess" (Kuhlemann 2018).

Kuhlemann argues that focusing exclusively on such risks is wrongheaded, and that scholars within ERS should broaden the threat horizon of interest to "unsexy risks" as well. She defines these as dangerous scenarios that "may also contain a low probability of an existential outcome, but what is most salient about them is the high probability of sub-existential outcomes." 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 *behaviorally*

[26] For example, Musk has said, "I will go [to Mars] if I can be assured that SpaceX would go on without me ... I've said I want to die on Mars, just not on impact" (quoted in Tiffany 2018). And Thiel has expressed a strong interest in transfusions with the blood of young people for rejuvenation purposes (Lieber 2019).

and attitudinally driven in the sense that their primary causes are "the procreative and livelihood-seeking behaviours constitutive of population growth and economic growth," these behaviors being "supported by attitudinal predispositions to oppose the kind of regulation of individual freedoms that could address the [risks] while curbing free riding" (Kuhlemann 2018). Examples include phenomena like "topsoil degradation and erosion, biodiversity loss, overfishing, freshwater scarcity, mass un- and under-employment, fiscal unsustainability, and … overpopulation" (Kuhlemann 2018).

This emphasis on unsexy risks, in what we can eponymously call the "Kuhlemann approach," is motivated in part by a rejection of the futurist perspective. Recall from section 2 that this coincides with the first paradigm, although its core ideas animate aspects of the second paradigm, too, and can be characterized 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 (Kuhlemann 2018). When one switches from the futurist to the normative perspective, the value-theoretic 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 Kuhlemann approach may provide an explanation of why some risks receive more scholarly attention than others, but one could also contend that the methodology of the first and second paradigms, informed largely by the etiological approach, is fundamentally flawed. Hin-Yan Liu, Kristian Cedervall Lauta, and Matthijs Maas (2018) defend this very thesis. They begin by noting that the foci of ERS research "are frequently characterised by relatively singular origin events and concrete pathways of harm which directly jeopardize the survival of humanity, or undercut its potential for long-term technological progress." Put differently, ERS has been narrowly fixated on the identification and mitigation of existential *hazards*, or the "discrete causes" and "external source[s] of peril" that could trigger an existential catastrophe. But this misses two other crucial factors: *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" (Liu et al. 2018). 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). Let's call this the "HVE approach."

More technically, ERS scholarship has habitually adopted the overly simple formula: "Existential risk = *existential hazard*." The first change to make here is adding variables that correspond to vulnerabilities and exposures, thus yielding a formula with three multiplicands: "Existential risk = existential hazard x vulnerability x exposure." But this too is problematic, since it suggests that the "existential" part of "existential risk" is associated with *and only with* the hazard component, which need not be the case. For example, Liu et al. (2018) 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." This yields two additional variants formulas: "Existential risk = hazard x existential vulnerability x exposure" and "Existential risk = hazard x vulnerability x existential exposure."

To illustrate, consider the following hypothetical scenario, which Kuhlemann would classify as "unsexy": ecological destruction, overuse of pesticides, aridification, deforestation, *and so on*, cause a major loss of pollinators. Without pollinators, flowering plants stop reproducing, followed by shrubs and trees. This causes land vertebrates and birds to perish, while bacteria and fungi populations explode

through metabolizing the deceased plants and animals. The soil further deteriorates, but humanity manages to survive on "wind-pollinated grains and marine fishing" (although the abundance of wild seafood is projected to disappear by 2048[27]). Consequently, the global human population tanks and "the wars for control of the dwindling resources, the suffering, and the tumultuous decline to dark-age barbarism" causes civilization to collapse and humanity to flirt with extinction (Wilson 2006). This is a realistic possibility in our future, but understanding the circuitous, criss-crossing routes from global stability to worldwide collapse requires examining the nexus between hazards, vulnerabilities, and exposures, all three of which conspired together to bring about this ruinous outcome.

Two important lessons emerge from the HVE approach. The first is that focusing only on existential hazards could actually *increase* the overall threat. The reason is that mitigating existential hazards could produce the false impression that we are safer than we are; the safety is thus merely "symbolic." As Liu et al. (2018) 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.

The second lesson is hortatory: ERS must expand its menu of strategies to address all three categories of causal factors. This implies that (a) ERS should work to 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, technological, etc. issues.

Another important development within the third paradigm is the "Cambridge approach," so-named because it emerged from a collaboration of scholars based at the Centre for the Study of Existential Risk (CSER) in Cambridge. This approach was first articulated in a(nother) 2018 paper by Shahar Avin, Bonnie Wintle, Julius Weitzdörfer, Seán Ó hÉigeartaigh,William Sutherland, and Martin Rees. They begin by noting, once again, 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. (2018) propose a comprehensive framework that identifies three primary contributory factors for global catastrophe:

> (1) One or more *critical systems*, demarcated by "safety boundaries," that a potential threat could breach. The authors recognize seven critical systems, each of which depends on the system "below" it in a hierarchy: *sociotechnological*, *ecological*, *whole organism*, *anatomical*, *cellular*, *biogeochemical*, and *physical*. Within each system, they further identify numerous 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* includes "climate control," "food," "health," "resource extraction," "security," "shelter," and "utilities" (Avin et al. 2018). Understanding each critical system thus requires knowledge of the various components that distinctly comprise it.
> (2) One or more *global spread mechanisms* that enable the threat 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, does 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

[27] See Worm et al. 2006.

> identify relevant mechanisms (and means to manage or control them) as targets of study meriting further attention, and highlights interesting commonalities" (Avin et al. 2018). They identify three primary spread mechanisms: *natural global scale*, *anthropogenic networks*, and *replicators*. An example of the former would be "air-based dispersal," which could enable supervolcanoes, asteroids, comets, and urban firestorms (following a nuclear conflict) alike to blot out the sun, thus causing worldwide crop failures. Or consider the replicators category. This 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 constituent 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*.

My aim is not to recapitulate every important detail of the Cambridge approach, but to underline further the paradigmatic shift away from reductionistic analyses of isolated phenomenon that characterizes the first and second paradigms. Indeed, the "critical systems" element above was inspired, at least in part, by a 2014 paper co-authored by one of the leading scholars within the third paradigm, Seth Baum. In this paper, Baum and his co-author, Itsuki Handoh, integrate the influential "planetary boundaries" framework proposed by Rockström et al. (2009) with what they call the "global catastrophic risk" (GCR) paradigm, which yields a novel risk framework that they call "Boundary Risk for Humanity and Nature" (BRIHN). In brief, the BRIHN framework yields a more precise definition of GCR that doesn't, as past definitions have, rely on arbitrary numbers of deaths or losses in GDP. Rather, they define GCR as the risks "of crossing a large and damaging human system threshold," where "crossing such a threshold could involve abrupt and/or irreversible harms to the human system, possibly sending the human system into a completely different state. The new state could involve significantly diminished populations and levels of development, or even outright extinction" (Baum and Handoh 2014). The "BRIHN approach," as we can call it, thus constitutes an early attempt within ERS to redirect the spotlight of scholarly attention away from epistemically neat—to borrow Kuhlemann's (2018) term—scenarios and instead analyze how potential disaster scenarios could unfold from a more "systems theory" perspective.[28]

Yet another development is also worth noting. This resulted from a renewed interest in the various kinds of state and nonstate actors who would either willingly (terror) or accidentally (error) destroy the world if only the means were available.[29] In his 1996 book, Leslie considers a cluster of "risks from philosophy," as he idiosyncratically calls them, such as anti-natalism and negative utilitarianism. This attentiveness to ideology was lost with Bostrom's 2002 publication, which fixated—unsurprisingly

---

[28] Two additional papers worth noting were *also* published in 2018. First, [redacted]. And second, an article co-authored by a collection of scholars that the BBC (2019) dubbed the "Trajectories Group." Titled "Long-Term Trajectories of Human Civilization," it explored four possible future trajectories of civilization, 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." As previously alluded to, the present author believes that (3) would lead to (4), and (4) would almost certainly instantiate (2).

[29] So far as I know, the terror/error distinction originated from Rees 2003.

given transhumanism's obsession with technology—almost exclusively on what we can call *technogenic* rather than *agential* threats.[30] In recent years, though, ERS scholars like the present author have explicitly 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 nonstate actors like small groups and even single individuals. I have thus proposed the somewhat cumbersome 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" (author 2018a, forthcoming). 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* (author 2018a, 2018b). The point is that 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.

Finally, the third paradigm has pushed back against certain canonical ideas within the first and second paradigms. For example, there is growing dissatisfaction with the assumption that developing dangerous dual-use technologies is inevitable, an idea encapsulated by Bostrom's (2009) "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." Kurzweil (2005) similarly holds that the creation of GNR (genetics, nanotech, and robotics) technologies is inexorable. But is this true? If the "default outcome" of making a value-misaligned superintelligence is "doom" (Bostrom 2014), then why not impose a moratorium on such research? Another idea that has been scrutinized is that space colonization constitutes an "existential panacea" that will vastly decrease the probability of extinction. But as I have argued elsewhere, following the lead of Daniel Deudney (forthcoming), there are strong reasons for believing that venturing into space could have catastrophic consequences, likely causing an s-catastrophe (author 2018). There is no "Planet B," as the environmentalist slogan goes; if humanity destroys this pale blue dot, then it destroys itself.

Section 5: Conclusion

Systematic investigation of humanity's future from a secular perspective is disappointingly new in history—indeed, the field of future studies wasn't "founded" until H.G. Wells published his 1901 book *Anticipations*, followed a year later by a Royal Institution lecture titled "The Discovery of the Future" (see Wager 1983). The past almost two decades, though, have witnessed the formation of a new field of scientific and philosophical inquiry focused on "existential risks" and multiple complementary and competing paradigms. At present, the two dominant thought-traditions are EA longtermism, which traces its genealogy to the futurist model, and analyses of catastrophic risk from a more systems-theoretic perspective. Although there are incompatibilities between the two, my contention here is that, given the incipiency of ERS, both offer some valuable insights about how we should understand, classify, and study existential risks, as well as why one should care about the topic in the first place.[31] This being said, I find myself far more sympathetic with the third paradigm than the second, which I believe suffers from serious

[30] Nonetheless, Lord Martin Rees pays some attention to the threats posed by a "lone dissident or terrorist," "embittered loners and dissident groups," and "disaffected loners" (Rees 2003; see author 2017).

[31] In this sense, I advocate a pluralistic *meta-ERS* position that embraces difference rather than advocating for conformity, a position that mirrors the "context-dependent semantic pluralism" that I have defended with respect to the term "existential risk" (author 2019).

problems with its motivating value system and methodology, that is, total utilitarianism and expected value theory.[32]

How might ERS evolve further in the coming years or decades? Here I 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 anomalies like lethal heatwaves, megadroughts, coastal flooding, rising sea-levels, melting glaciers and polar ice caps, desertification, food supply disruptions, zombie pathogens, more infectious disease, biodiversity loss, species extinctions, state shifts in the global ecosystem, economic collapse, mass migrations, social upheaval, political instability, cultural and religious clashes, interstate and civil wars, terrorist attacks (see author 2017), *and so on*. As interest in the topic grows, more media outlets will cover the day's news and, in doing so, consult with experts who likely will have by this time stumbled upon the concept of *existential risk* and perhaps perused the corresponding literature, especially if the field successfully spreads its tentacles into other disciplines. Already, *Vox Media* has a vertical, *Future Perfect*, that is dedicated entirely to global catastrophic and existential risks, as well as other EA causes like global poverty and animal welfare. Mass movements like "Extinction Rebellion" and "Skolstrejk för Klimatet" ("school strike for climate") could also make human extinction and civilizational collapse increasingly visible to the public, thereby amplifying public interest.

Second, novel existential risk scenarios—which I have elsewhere termed "monsters" (author 2016, forthcoming)—could pop up on the threat horizon. Consider the fact that risks associated with nuclear war, engineered pandemics, superintelligence, and so on, would have been unimaginable to, say, Charles Darwin or Lord Kelvin in the nineteenth century. We may be in the same epistemic predicament with respect to any number of future threats; as I have previously noted, a textbook on existential risks written in 2060 might very well contain 10 times as many chapters than one written today, each chapter covering a different risk scenario (see author 2017). Even more, as Anders Sandberg, Jason Matheny, and Milan Ćirković observe, writing in 2008, supervolcanism "was discovered only in the last 25 years, [which suggests] that other natural hazards may remain unrecognized." Perhaps there are—I say somewhat facetiously—*super-earthquakes* that could cause mass extinctions but to date have eluded scientific sleuthing by geologists. This may sound cockamamie, yet while writing this paper scientists announced the discovery of a new type of seismic phenomenon that they dubbed "stormquakes." A stormquake occurs when storms create secondary waves that cause the seafloor to shake enough for detection by geological survey instruments. The point is that if more existential risk scenarios are either created by human activity ("ontological risk multiplication") or discovered by science ("epistemic risk multiplication"), then the ranks of ERS could further swell (see author, forthcoming).

Third, ERS has been dominated until quite recently, in large part because of the third paradigm, by white men. This has resulted in certain issues being foregrounded more or less than they otherwise would have been if the field had been more diverse in terms of ideology, race, gender, disability, and so on. For example, many marginalized peoples throughout the world do not have the luxury of engaging in armchair speculation about the supposed astronomical value of the far future once our posthuman descendants colonize the universe, subjugate nature, and maximize economic productivity. Similarly, it may appear profoundly callous from certain perspectives to assert that it would be *just as good* to reduce the probability of existential risk by 0.00000000000000000001 percent if there's a 1 percent chance of

---

[32] Consider that total utilitarianism engenders the Repugnant Conclusion, so-named because Parfit (1984) found it completely unacceptable. Yet, as Hilary Greaves (2017) observes, various "impossibility theorems" show that there can be no Parfitian "Theory X" that would, by definition, avoid the Repugnant Conclusion, avoid the Sadistic Conclusion, and respect non-anti-Egalitarianism. Greaves then suggests that the Repugnant Conclusion may only be repugnant because of "*untutored intuition*" arising from "distorting biases."

10^{54} currently non-existent, merely possible future people than to save 1 billion living, breathing human beings alive today, especially when Western colonialism, imperialism, military intervention, political meddling, and environmental degradation are responsible for the dismal plight of so many in the "Global South." In fact, expected value calculations like the one just above led Beckstead (2013) to conclude that, since privileged rich people in the Western world (also mostly white) are in a better position to shape the far future, resources ought to be allocated to *them* rather than towards impoverished, starving people elsewhere. In his words,

> saving lives in poor countries may have significantly smaller ripple effects than saving and improving lives in rich countries. Why? Richer countries have substantially more innovation, and their workers are much more economically productive. By ordinary standards—at least by ordinary enlightened humanitarian standards—saving and improving lives in rich countries is about equally as important as saving and improving lives in poor countries, provided lives are improved by roughly comparable amounts. But it now seems more plausible to me that saving a life in a rich country is substantially more important than saving a life in a poor country, other things being equal (Beckstead 2013).[33]

Consider here Kuhlemann's (2018) distinction between "future people" and "hypothetical people." The former refers to human beings who are "assumed to be alive at a future time" whereas the latter refers to human beings who "could be caused to exist at a future time, but until such time are merely imaginary." From the normative perspective that she defends, what matters is "safeguarding the interest of and preventing harm to *actual* human beings, whether alive now or in the future," while the futurist aim, advocated by Beckstead, is "safeguarding the potential utility to be held or derived by all human-descent lives that already exist or could exist in the future" (Kuhlemann 2018).

The crux is that there may be a direct connection between white male privilege and the futurist perspective, which tends to dismiss the suffering of people today as trivial *when compared to* the well-being or value that could come to exist in the distant future. In fact, a meta-analytic re-analysis 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" (Friesdorf et al. 2015). This suggests that a more gender diverse field might drift away from methodological habits like plugging numbers into decision-theoretic algorithms and calculating the relative *unimportance* of saving 1 billion contemporary lives. Perhaps there are ways of thinking about these issues that are not even conceivable to contemporary ERS scholars—especially those whose perspectives have been molded by white and/or male privilege—but that could coalesce in the future, if ERS continues to diversify.

Humanity has only recently become aware of its own mortality, and only in the past few decades begun to systematically study the existential risks that could end our collective story forever. This paper has attempted to sketch out the historical evolution of this field from roughly 2002 until the present, arguing that two overlapping but distinct ways of thinking about our existential predicament currently dominate ERS research. My 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: existential risk—whatever that means exactly.

Acknowledgements: Thanks to Matthijs Maas, Dan Elton, David Pearce, Alexey Turchin, and Azita Chellappoo for many insightful comments on a draft.

References:

[33] Some might simply interpret this as a knock-down argument against utilitarianism.

Lieber, Chavie. 2019. The FDA Says Buying Young People's Blood Won't Stop You From Aging. *Vox*. https://www.vox.com/the-goods/2019/2/19/18232259/fda-young-blood-transfusion-safety-concerns.

Tiffany, Kaitlyn. 2018. Elon Musk Thinks He'll Die on Mars. *Vox*. https://www.vox.com/the-goods/2018/11/2/18053824/elon-musk-death-mars-spacex-kara-swisher-interview.

Bostrom, Nick. 2002. Existential Risks: Analyzing Human Extinction Scenarios and Related Phenomena. *Journal of Evolution and Technology*. 9(1).

Worm, Boris, Edward Barbier, Nicola Beumont, et al. 2006. "Impacts of biodiversity loss on ocean ecosystem services." Science 314 (3 November 2006): 787-790.