Is Existential Risk a Useless Category? Could the Concept Be Dangerous?

Abstract: This paper offers a number of reasons for why the Bostromian notion of *existential risk* is useless. On the one hand, it is predicated on a highly idiosyncratic techno-utopian vision of the future that few would find appealing. On the other, the "worst-case outcomes" for humanity group together the atrocious to the benign. What matters, on Bostrom's view, is not human extinction *per se*, but any event that would permanently prevent current or future people from attaining technological Utopia. I then consider the question of whether the Bostromian paradigm could be dangerous. My answer is affirmative: this perspective combines utopianism and utilitarianism. Historically, this has proven to be a highly combustible mix. When the ends justify the means, and when the end is paradise, then groups or individuals may feel justified in contravening any number of moral constraints on human behavior, including those that proscribe violent actions. Although I believe that studying low-probability, high-impact risks is extremely important, I urge scholars to abandon the Bostromian concept of *existential risk*.

1. Introduction

The past two decades has seen the emergence of a novel scholarly field known as "Existential Risk Studies" (ERS). The aim of the field is threefold: (i) delineate the etiology of existential risks; (ii) devise effective strategies for mitigating existential risks; and (iii) understand the axiological/ethical implications of existential risks. Although the field remains relatively small, several institutes based at elite universities, such as Oxford and Cambridge, have been founded to study the topic. The corresponding research literature has expanded considerably, es-

pecially in the past ten years. More recently, leading figures in the "Effective Altruism" (EA) movement, such as Toby Ord and William MacAskill, have taken a strong interest in existential risk, arguing that ensuring the continued survival and flourishing of humanity ought to constitute one of the top three charitable causes, if not the overriding cause. Numerous popular media platforms have also introduced the idea of existential risks to the general public (e.g., see Khatchadourian 2015). Most salient is *Vox*'s vertical "Future Perfect," which is dedicated to reporting on issues relevant to EA concerns and the long-term future of humanity.

But there has also been some criticism of the ERS research program. Perhaps most notably, Steven Pinker has argued that some of the potential sources of existential risk are too speculative to warrant serious consideration. Such "Frankensteinian fantasies," he argues, distract from the two genuine dangers to human well-being on a global scale: anthropogenic climate change and thermonuclear conflict. As Pinker (2018) writes in *Enlightenment Now*, "sowing fear about hypothetical disasters, far from safeguarding the future of humanity, can endanger it." He consequently contends that *existential risk* is a "useless category" (Kupferschmidt 2018). While I strongly disagree with Pinker's (generally misinformed) dismissal of threats associated with synthetic biology, nanotechnology, artificial superintelligence, and other emerging technologies (see Author 2018a for a detailed critique), I do concur that the canonical concept of *existential risk* is useless (although for different reasons than Pinker). Even more, it could nontrivially increase the likelihood of global catastrophes and mass atrocities, that is, if it were to become sufficiently influential within certain intellectual communities and among our political leaders. The present paper offers myriad reasons for accepting these alarming conclusions. I will proceed as follows: section 2 outlines an argument for why "existential risk" in the *particular sense* of Nick Bostrom's categorization (2002, 2013) is useless. Hence, this is not a critique of ERS in general,

but one paradigm within ERS in particular (the dominant paradigm). Section 3 explains why the theoretical framework in which the "Bostromian" concept of *existential risk* is ensconced poses clear and present dangers to humanity. And section 3 concludes with a short recap of the main points.

2. What Are Existential Risks?

2.1 Theoretical Framework: There are two main reasons that the concept of *existential risk* constitutes a useless category. First, the most prominent definition among existential risk scholars is based on a techno-utopian vision of the future that few would, upon reflection, accept. Second, it lumps together under the same umbrella scenarios that range from the unimaginably atrocious to the utterly benign. To understand these assertions, we need a clear understanding of the factors that shaped the standard conception of existential risk. There are, to be clear, multiple definitions of the term in the scholarly literature. For example, some scholars equate the *existential* with the *extinctional*, and thus define "existential risks" as "risks to human survival."[1] Others adopt a more promiscuous semantics and include global catastrophes like civilizational collapse scenarios within the term's extension (see Author 2019). But the original formulation of "existential risk," which remains the canonical definition within ERS, comes from two papers from Bostrom, published more than a decade apart: "Existential Risk: Analyzing Human Extinction Scenarios and Related Hazards" (2002) and "Existential Risk as Global Priority" (2013).

Bostrom provides several variants of two general formulations: a *lexicographic definition* and a *typological definition* (Author 2019). These are not equivalent (although it appears that they were intended to be) and, as I have discussed before, the latter is based on a typology of

risks in general that commits multiple "category mistakes"; thus I will not discuss it here. In his 2002 paper, Bostrom offers the first lexicographic definition of existential risk as "one where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential." In his 2013 paper, he offers a theoretically more refined definition: "one that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development." The question, then, is what "premature extinction" and "desirable future development" mean. Bostrom's answer pertains to what he calls "technological maturity," which refers to "the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved." Hence, on this view, an existential risk is any event that would prevent "humanity" (which includes our posthuman descendants) from attaining a stable state in which we have subjugated nature and maximized economic productivity as close to the absolute physical limits as possible. Why does this Baconian, capitalistic *telos* matter? Because of two underlying views that Bostrom champions, namely, *transhumanism* and *total utilitarianism*. Taking these in turn:

2.1.1 Transhumanism. Transhumanism is a normative ideology whose "core value" is "having the opportunity to explore the transhuman and posthuman realms" (Bostrom 2005). According to Bostrom, a "basic condition" for realizing this core value is "technological progress."[2] The idea is that posthumans could have lives that are unfathomably more "fulfilling" (Bostrom 2013) than the lives of current humans; since more fulfilling lives are better than less fulfilling lives, and since technological maturity is necessary for maximally enabling current or future people to explore posthuman modes of being, the full realization of the core value of transhumanism requires the attainment of technological maturity. What could such lives be like?

Bostrom offers a glimpse in his quasi-poetic "Letter from Utopia" (2008/2020). This is composed by a fictional posthuman in techno-utopia to contemporary people; hence, it is addressed "Dear Human" and signed "Your Possible Future Self." The posthuman asks, "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." He or she, perhaps living ecstatically inside a high-resolution computer simulation (see below), proceeds:

> 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. Quite some marvelous edifices build up over a thousand years by the efforts of homunculi, just as the humble polyps in time amass a coral reef. And I've seen the shoals of biography fishes, each one a life story, scintillate under heaving ocean waters.

To which the author adds:

> 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!

The author then writes that "to reach Utopia, you must discover the means to three fundamental transformations," namely, (i) *securing life*, i.e., becoming technologically immortal; (ii) *expanding cognition*, i.e., becoming superintelligent, as "it is in the spacetime of awareness that Utopia will exist"; and (iii) *elevate well-being*, i.e., maximize pleasure: "a few grains of this magic ingredient are worth more than a king's treasure" (Bostrom 2008/2020). Other transhumanists have identified different fundamental transformations. For example, Ray Kurzweil (2006) argues that techno-utopia will be ushered in by a history-rupturing event called the "technological Singularity," which

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

In the rigid secular eschatological narrative outlined by Kurzweil (2005), which consists of six basic epochs in cosmic history, the Singularity will occur in 2045. Bostrom does not specify a date, but once argued that "we will have superhuman artificial intelligence within the first third of the next century" (Bostrom 1997).[3] The creation of machine superintelligence may be integral for what Bostrom (and other transhumanists) call "paradise-engineering," which involves the elimination of all human suffering. Others have taken this idea a step further and argued that we

should use advanced technologies to reengineer the entire biosphere to eliminate all sentient suffering. For example, David Pearce, who co-founded the World Transhumanist Association in 1998 with Bostrom, advocates for what he calls the "abolitionist project" in his book *The Hedonistic Imperative*.

While these ideas may sound visionary (and I suspect the authors have seen them as such), history is overflowing with utopian fantasies of a paradisiacal world—or otherworld—to come, including "secular" fantasies like those associated with Marxism-Leninism and Nazism (see Author 2016a, 2016b). Perhaps the single most influential text in all of history (Flannery 2016) is deeply utopian in its prophetic proclamations: the Book of Revelation, the final book of the New Testament. Indeed, its description of paradise is remarkably similar to those found in "Letter from Utopia," Kurzweil's tomes, and the writings of other transhumanists. For example, Revelation 21:4 declares that God "'will wipe every tear from their eyes. There will be no more death' or mourning or crying or pain, for the old order of things has passed away." In the secular-transhumanist account, there is no God, but machine superintelligence plays a similarly divine role. For example, it is widely thought that superintelligence will either cause our extinction or bring about a techno-utopian world: "whether we succeed or fail—it is probably the last challenge we will ever face," Bostrom (2014) asserts, because "one might believe that … while the creation of superintelligence will pose grave risks, once that creation and its immediate aftermath have been survived, the new civilization would have vastly improved survival prospects since it would be guided by superintelligent foresight and planning" (Bostrom 2009).

2.1.2 Totalism. As for total utilitarianism (or "totalism"): this identifies value—specifically, intrinsic value—with well-being, defined in hedonistic, desire-satisfactionist, or objective list theoretic terms (see Parfit 1984). From the Sidgwickian "point of view of the universe," a state of

affairs that instantiates more intrinsic value is better than one that instantiates less, and a state of affairs that instantiates a maximal amount of value would be best. Since people are the substrate, or "containers" (Rawls 1974), of intrinsic value, total utilitarianism demands that we maximize the total number of people who come to exist with "worthwhile" or "happy" lives, meaning that they contain a net positive amount of well-being. Put differently, people are not intrinsically valuable, but only *instrumentally valuable*: we are means to an end, the end being the net value or well-being that we introduce into the universe, which is good from the universe's point of view. According to Bostrom (2003) and others (Cirkovic 2002), the total number of people who could come to exist within our future light cone is *astronomical*, especially if our descendants colonize as much of the universe as possible, as quickly as possible, converting whole exoplanets into computronium on which massive simulations crowded with "happy" people. As he writes, "many more orders of magnitudes of human-like beings could exist if we countenance digital implementations of minds—as we should" (Bostrom 2014). This could result in some 10^{38} simulated people per century in our supercluster alone (Bostrom 2003), although Bostrom (2014) offers a much larger number later on: "assuming that the observable universe is void of extraterrestrial civilizations, then what hangs in the balance is at least 10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 human lives (though the true number is probably larger)." He adds that "we might get additional orders of magnitude" of operations per second "if we make extensive use of reversible computation, if we perform the computations at colder temperatures (by waiting until the universe has cooled further), or if we make use of additional sources of energy (such as dark matter)."

Bostrom imagines these simulated people living in simulated universes with "rich and happy lives while interacting with one another in virtual environments" (Bostrom 2014). What

might they actually be like on a quotidian level? It is impossible to say for sure; if these trillions and trillions and trillions (and trillions and trillions) of people are total utilitarians, perhaps they would pursue activities with a sense of moral duty knowing that they are contributing to the total amount of net value in the universe, which is good from the impersonal perspective of the universal gaze (Srinivasan 2015). What is for sure is that, given how many possible people could become actual people, "if we represent all the happiness experienced during one entire such life with a single teardrop of joy, then the happiness of these souls could fill and refill the Earth's oceans every second, and keep doing so for a hundred billion billion millennia" (Bostrom 2014). This of course sounds reminiscent of the Revelation passage mentioned above, and indeed the motif of *oceans of tears* recurs frequently throughout religious texts. For example, Jonathan Edwards (1733) once wrote that "to go to heaven, fully to enjoy God, is infinitely better than the most pleasant accommodations here. Better than fathers and mothers, husbands, wives or children, or the company of any, or all earthly friends. … These are but drops; but God is the ocean." And John Allen (1802) exhorted: "Come, Christian, launch forth into the ocean of joy." An interminable list of other examples could be adduced.[4] The point is that, for Bostrom, an existential risk is any event that would prevent the stable realization of a techno-utopian state of affairs in which astronomical numbers of people—perhaps most being simulated—are able to explore the posthuman realm, and thus live far richer and more fulfilling lives than we currently have. But to realize this state of affairs, we must fully subjugate all natural processes and maximize economic productivity to the physical limits, which is why technological maturity is the central teleological component of Bostrom's (2013) definition of "existential risk."

When understood thusly, I am doubtful that many people would find this version of techno-utopia, or "Utopia," appealing. Indeed, utilitarianism itself is a minority view among academ-

ic philosophers (see Bourget and Chalmers 2014), in part because of its *highly* counterintuitive implications with respect to certain ethical questions. For example, it implies that a doctor should kill a healthy person and harvest her organs to save five others, and since people are mere instrumental containers, there would be nothing objectionable to instantaneously replacing every living person at T1 with a different person at the next moment T2 if the T2 people instantiate the same total amount of value, in aggregate, as the T1 people instantiated, in aggregate (see Knutsson 2019). You do not matter because of you, but because of what you contribute to the total net sum of well-being in the universe. Furthermore, the utopian aspects of transhumanism may appear quixotic and insufficiently thought through. There have to date been few good critiques of the transhumanist project, but this is not, I believe, because there are no good criticisms to level at this worldview. Of note is Yuval Noah Harari's (2015) worry that radical human enhancement would result in an elite class of enhanced humans and an impoverished class of unenhanced humans. To quote him at length:

> Some people will remain both indispensable and undecipherable, but they will constitute a small and privileged elite of upgraded humans. These superhumans will enjoy unheard-of abilities and unprecedented creativity, which will allow them to go on making many of the most important decisions in the world. They will perform crucial services for the system, while the system could not understand and manage them. However, most humans will not be upgraded, and they will consequently become an inferior caste, dominated by both computer algorithms and the new superhumans. … The great human projects of the twentieth century—overcoming famine, plague and war—aimed to safeguard a universal norm of abundance, health and peace for all people without exception. The new projects

> of the twenty-first century—gaining immortality, bliss and divinity—also hope to serve the whole of humankind. However, because these projects aim at surpassing rather than safeguarding the norm, they may well result in the creation of a new superhuman caste that will abandon its liberal roots and treat normal humans no better than nineteenth-century Europeans treated Africans.

This strikes me as extremely plausible: the socioeconomically advantaged will be motivated to ensure status quo power dynamics by impeding lower-class access to enhancements. There is nothing to lose and everything to gain by doing this, from a self-interested perspective (which is precisely the perspective that capitalist economies encourage); I am not even sure why the enhanced elite would choose not to slaughter the lowly *Homo sapiens* remaining (consider our treatment of other animals, especially "vermin" that "cause a nuisance"). There are also strong reasons for believing that colonizing space could foment catastrophic conflict between technologically enhanced posthuman populations, for reasons pertaining to game theory and neorealism in international relations theory (see Author 2018b; Deudney forthcoming). Although Milan Cirkovic (2019) has responded to some of these arguments, the main thrust of his objections are rooted in utopian-transhumanist thinking: postbiological creatures will magically be peaceable rather than bellicose, and consequently they will manage to effectively extricate themselves from the security dilemma and Hobbesian traps that both myself and Deudney identify as conflict-generators within the anarchic cosmopolitical realm. In a phrase, it is easy to imagine dystopian worlds in which no one would want to live; it is equally as difficult to imagine a utopian paradise in which anyone would actually want to be.[5]

These features make Bostrom's notion of existential risk quite useless *unless* one also accepts the quasi-religious ideologies of transhumanism and total utilitarianism, as well as the Baconian/capitalistic desiderata of subjugating nature and ramping up economic productivity to the absolute limits. On the Bostromian view, the universe is a vast reservoir of negentropy just waiting to be exploited, and every second that we fail to spread into the cosmos and exploit nature is a second during which astronomical numbers of possible happy people/value-containers (most living in computer simulations) fail to come into existence, which the universe sees as morally bad. But if one rejects transhumanism, total utilitarianism, the Baconian imperative to gain dominion over nature, and/or the capitalistic urge to maximize productivity, then one will reject Bostrom's definition of "existential risk" as a bit of ideological propaganda. At first glance this may appear reasonable, but upon closer examination it entails a quite radical view about what the ultimate goals of our technological civilization ought to be.

2.2 Disparate Scenarios. Another objectionable aspect of this definition is that it fails to differentiate between scenarios that range from, as stated above, the unimaginably atrocious to the utterly benign. Consider, for example, the following scenario S1: imagine that a group of omnicidal agents (see Author 2018c, 2018d) fabricate a "deep fake" of the President of the United States announcing that US intelligence agencies have determined with "tremendous certainty" that China and Russia are about to launch a preemptive nuclear strike against the US mainland. The military's top generals have all advised, "like you wouldn't believe," an immediate counter-strike to neutralize this existential threat. The group hacks into the President's twitter account and posts this video, which quickly goes viral around the world, causing panic and alarm among both foreign citizens and governments. The US government immediately responds to the incident, calling it a "hack," but messages are mixed and confusing: some deem this response a piece of

"fake news" from the "Deep State" that is intended to pacify the masses to maintain law and order during the strike. Meanwhile, China and Russia launch a barrage of thermonuclear missiles at US urban centers. These initiate firestorms that pollute the stratosphere with sunlight-blocking soot, which diffuses around the entire globe causing a nuclear winter. Although China and Russia did not intend this outcome, the models of how much soot would be produced by US cities were wrong (call this "Castle Boo-hoo"). Global agriculture implodes, causing the excruciatingly slow death of huge numbers of people from starvation; malnutrition enables the outbreak of new infectious diseases that result in pneumonia, hemorrhaging, and neurological disorders. Societies around the world collapse and Hobbesian anarchy becomes the new abnormal. Localized tribes form around shared ideologies; those with fortified compounds, bunkers, arsenals of assault weapons, and large stashes of ammunition kill off other tribes to seize control of scarce resources. Constant fear of attack leads groups to engage in first strikes against each other, thereby increasing the death toll. Yet over the course of nearly a decade of pitch-black skies at noon, food supplies dwindle such that even the best-off "doomsday prepper" communities die out. The human species goes extinct.

Compare this with another scenario S2: technological development proceeds from the time of this writing (in 2020) for another decade. Cures for pathologies like Alzheimer's, diabetes, and heart disease are discovered. New strategies for preventing large-scale outbreaks of infectious disease are developed, and life expectancy around the world increases to 95 years old. The human population stabilizes at around 8 billion people, and advancements in food production enable nearly everyone around the world—in both the Global North and Global South—to meet their daily nutritional needs. Furthermore, advancements in sustainable technology enables civilization to remain within its carbon budget, thus avoiding climatic warming of more than 1.5

degrees Celsius above pre-industrial levels. Global travel becomes easy for nearly everyone, democracy spreads, violence continues to decline, and moral progress leads to increasing tolerance around the world for LGBTQ people, equality for women, liberation for animals, and so on (see Pinker 2011). But at the end of this decade, technological progress stalls permanently: the conditions realized at the end of the decade are the conditions that hold for the next 1 billion years, at which point Earth becomes uninhabitable due to the sun's growing luminosity. Nonetheless, many trillions and trillions of humans will come to exist in these conditions, with more opportunities for self-actualization than ever before.

Allow me to ask the rhetorical question: Which of these two scenarios is worse? By how much is one worse than the other? Clearly, S1 is worse than S2, by a lot. Yet according to Bostrom's lexicographic definition, *both* S1 and S2 are existential risks—they are co-categorical disaster scenarios, undifferentiated in their ontological status as the worst-possible outcomes for humanity. After all, in both cases, astronomical numbers of future posthumans (many or most living in computronium-substrate simulations) throughout the visible universe will never come into existence. This is based on the assumption that, as Peter Singer (2015) puts it, summarizing Bostrom's view, "the value lost when an existing person dies is no greater than the value lost when a child is not conceived, [given that] the quality and duration of the lives are the same." That is to say, if one assumes that, as Bostrom asks us to, "holding the quality and duration of a life constant, its value does not depend on when it occurs or on whether it already exists or is yet to be brought into existence as a result of future events and choices," then the numerical difference between S1 and S2 is that (a) S2 entails the loss of astronomical amounts of future value whereas (b) S1 entails the loss of astronomical amounts of future value plus 8 billion people alive at the time of the nuclear catastrophe. Since 8,000,000,000 actual people is *tiny* compared

to (probably more than)

10,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000 possible value-containers, it follows that the *axiological* difference between S1 and S2 is negligible. Just consider Bostrom's (2002) claim about some of the worst disasters and atrocities in human history, such as "Chernobyl, Bhopal, volcano eruptions, earthquakes, draughts, World War I, World War II, epidemics of influenza, smallpox, black plague, and AIDS." For Bostrom,

> these types of disasters have occurred many times and our cultural attitudes towards risk have been shaped by trial-and-error in managing such hazards. But tragic as such events are to the people immediately affected, in the big picture of things—from the perspective of humankind as a whole—even the worst of these catastrophes are mere ripples on the surface of the great sea of life. They haven't significantly affected the total amount of human suffering or happiness or determined the long-term fate of our species.

Since these disasters haven't "significantly affected the total amount" of value that will exist in our future light cone, they do not constitute existential risks; and since they do not constitute existential risks, humanity should not prioritize the avoidance of similar future events—great wars, epidemics, the Holocaust the Rwandan genocide. Rather, as Bostrom (2003) writes, "for standard utilitarians, priority number one, two, three and four should … be to reduce existential risk," where priority number five should (presumably) be to colonize the visible universe as quickly as possible: on his calculations, about 10^{29} potential human lives are lost every second of delayed colonization. Or as Bostrom (2013) later puts the point,

> unrestricted altruism is not so common that we can afford to fritter it away on a plethora of feel-good projects of suboptimal efficacy. If benefiting humanity by increasing existential safety achieves expected good on a scale many orders of magnitude greater than that of alternative contributions [which it does given the equivalence of someone dying and someone not being born], we would do well to focus on this most efficient philanthropy.

Singer (2015) provides two examples of causes that would, on Bostrom's view, constitute "feel-good projects" to which we should not "fritter away" resources, namely, "helping people in extreme poverty [and] reducing the suffering of animals."[6] These are both, by many accounts, ongoing global catastrophes. Yet Bostrom argues that it would be foolish to identify the causes of mitigating them as being among our top five global priorities. As he puts it, our limited resources for doing good in the world would be wasted if they were directed "with excessive preponderance" mainly "at global catastrophic risks that involve little or no existential risk" (Bostrom 2013). On the standard semantics, all existential risks are global catastrophic risks, but not all global catastrophic risks are existential risks. Thus, there are global catastrophic risks that would not permanently prevent humanity from realizing a techno-utopian world, and these mere "ripples" must not distract us from what really matters.

However, if one does *not* value currently non-existent, possibly never-existent people *as much as* people who do currently exist, then one will find this perspective obscene. Currently existing people are (to lean on familiar cliches) real, living, breathing, actual human beings full of hopes, dreams, fears, loves, and other conscious experiences. Real people can bleed; imaginary people cannot. On this anti-Bostromian (so-called "person-affecting") view, we may indeed wish to care about the class of people who do not yet exist but will independently of our actions

(Karin Kuhlemann calls these "future people"[7]), but it would be deeply misguided to assert, as Bostrom does, that *the worst aspect of human extinction by far* would not be the 8 billion conscious beings who would die but the 10^{58} non-conscious non-beings who would never be born. This yields a second reason for why "existential risk" is useless: it identifies, on dubious grounds, scenarios that clearly do not constitute worst-case outcomes for humanity (S2) with scenarios that clearly do constitute such outcomes (S1). A theoretical framework that cannot distinguish between S1 and S2 does not have any practical applications for determining our collective global priorities. Hence, scholars should abandon the *concept*, even if they decide to retain the *term*.

3. Is the Concept of Existential Risk Dangerous?

3.1 Utopianism and Utilitarianism. A concept being useless need not imply that it is also dangerous; there are many concepts, such as *overpopulation on Mars*, that are "useless" but harmless. In this section, I will argue that *existential risk*, as defined above, is not one of these concepts. The central reason is that utopianism and utilitarianism are highly combustible when combined: one prescribes moral choices that conduce to "the greater good," while the other identifies "the greater good" as *paradise*. To my knowledge, Bostromianism is the first ideology to *explicitly* combine these two elements; I am unfamiliar with any total utilitarians since the late-eighteenth century (when utilitarianism was systematically formulated) who embraced techno-utopian visions like those depicted in "Letter from Utopia."[8] Yet history abounds with ideologies that wove together utopian fantasies with utilitarian modes of moral calculation—and in many cases the result was, at one point or another in the lifetime of the ideology, catastrophic violence.

As mentioned above, both Marxism-Leninism and Nazism were motivated by eschatological worldviews according to which the end of history (as we know it) marks the beginning of utopia (or heaven on Earth). In the case of Marxism-Leninism, utopia is the realization of a "pure" communist world state, and class struggle, which could (and did) turn violent, is the means for reaching this climactic end. With respect to Nazism, Hitler preached of a "Thousand Year Reich" or "millennium of perfection," and "appealed to a higher law, to a mission decreed by fate" in arguing for bringing about this perfection (Chirot and McCauley 2006). Furthermore, perhaps the second most devastating conflict in human history, the Taiping Rebellion (1850-1864), was driven by the syncretistic utopianism of Hong Xiuquan, who led the Taiping Heavenly Kingdom against the Qing dynasty. The point is that when what is at stake is paradise, and if one holds that certain ends can justify certain means, then those who seriously believe the promise of paradise are liable to pursue whatever means are necessary to reach this end, even if these means violate moral constraints against, for example, harming others. As Pinker (2011) puts this point,

> utopian ideologies invite genocide for two reasons. One is that they set up a pernicious utilitarian calculus. In a utopia, everyone is happy forever, so its moral value is infinite. Most of us agree that it is ethically permissible to divert a runaway trolley that threatens to kill five people onto a side track where it would kill only one. But suppose it were a hundred million lives one could save by diverting the trolley, or a billion, or—projecting into the indefinite future—infinitely many. How many people would it be permissible to sacrifice to attain that infinite good? A few million can seem like a pretty good bargain.

The "moral value" that could come to exist on Bostrom's view is not infinite, but it is astronomical. It is, indeed, so astronomical that disasters like World War II and the Holocaust are mere blips—"ripples"—in the grand scheme of things. Avoiding such disasters is not even among our top four (or five) priorities, since such events do not pose permanent barriers between humanity today and Utopia tomorrow. If someone—a secular extremist—were to take this line of reasoning seriously, she could justify to herself a wide range of atrocities for the sake of "paradise-engineering." For example, consider the following scenario from Olle Häggström (2016); quoting him at length:

> Recall … Bostrom's conclusion about how reducing the probability of existential catastrophe by even a minuscule amount can be more important than saving the lives of a million people. While it is hard to find any flaw in his reasoning leading up to the conclusion [note: the present author objects], and while if the discussion remains sufficiently abstract I am inclined to accept it as correct, I feel extremely uneasy about the prospect that it might become recognized among politicians and decision-makers as a guide to policy worth taking literally. It is simply too reminiscent of the old saying "If you want to make an omelet, you must be willing to break a few eggs," which has typically been used to explain that a bit of genocide or so might be a good thing, if it can contribute to the goal of creating a future utopia. Imagine a situation where the head of the CIA explains to the US president that they have credible evidence that somewhere in Germany, there is a lunatic who is working on a doomsday weapon and intends to use it to wipe out humanity, and that this lunatic has a one-in-a-million chance of succeeding. They have no further information on the identity or whereabouts of this lunatic. If the president has taken

> Bostrom's argument to heart, and if he knows how to do the arithmetic, he may conclude that it is worthwhile conducting a full-scale nuclear assault on Germany to kill every single person within its borders.

Häggström offers several reasons why this scenario might not occur. For example, he suggests that "the annihilation of Germany would be bad for international political stability and increase existential risk from global nuclear war by more than one in a million." But he adds that we should wonder "whether we can trust that our world leaders understand [such] points." Ultimately, Häggström abandons total utilitarianism and embraces an absolutist deontological constraint according to which "there are things *that you simply cannot do*, no matter how much future value is at stake!" But not everyone would follow this lead, especially when assessing the situation from the point of view of the universe; one might claim that, paraphrasing Bostrom, as tragic as this event would be to the people immediately affected, in the big picture of things—from the perspective of humankind as a whole—it wouldn't significantly affect the total amount of human suffering or happiness or determine the long-term fate of our species, except to ensure that we continue to exist (thereby making it possible to colonize the universe, simulate vast numbers of people on exoplanetary computers, and so on). Or consider another example, which I will call "Buridan's altruist," on the model of "Buridan's ass."

> *Buridan's Altruist*: Imagine someone S sitting in front of two buttons. If S pushes the first button, 1 billion actual human beings will be prevented (somehow) from being executed by electrocution. If S pushes the second, S will reduce the probability of not reaching and sustaining techno-utopia by a barely noticeable amount. Which button should S push?

According to Bostrom, S should be paralyzed between these two options, since each is (*ex hypothesi*) equivalent in (expected) moral value. Now consider a variant, which I will call "Bostrom's altruist":

> *Bostrom's Altruist*: Imagine the same scenario as above, except that if S pushes the second button, S will reduce the probability of not reaching and sustaining techno-utopia by *slightly more* than a barely noticeable amount. Which button should S push?

The right moral choice is obvious, on the Bostromian view: S should push the second button. As Bostrom (2012) makes this very point: if there are 10^{54} people who could come to exist within our future light cone and this has "a mere 1% chance of being correct," then "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." These inscrutable numbers make the choice to push the second button not merely obvious; they imply that S *not* sacrificing billions and billions of people would make S a moral monster. This is a hypothetical and idealized case, but it is not difficult to imagine—as Häggström does—an analogous real-world situation. If the person facing such a moral choice were to take seriously Bostrom's vision of Utopia, of trillions and trillions and trillions (and trillions and trillions) of possible people living in computer simulations cluttering our future light cone, the result could be catastrophe. We should therefore hope that the Bostromian ideology does not become so widespread or well-known as to influence whoever ends up sitting in front of these two buttons.

This is one way that the concept of *existential risk* could be dangerous—even *catastrophic*, like the ideologies of Marxism-Leninism, Nazism, and the Taiping Heavenly Kingdom. I do not believe that this is hyperbole: many or all of the ingredients needed for mass atrocities are present in the Bostromian paradigm. Here one may observe, perhaps as an objection, that this paradigm is not *revolutionary* (or *apocalyptic*) in the way that Marxism-Leninism, Nazism, and the Taiping Heavenly Kingdom were. Bostrom's eschatological narratives do not posit that there *must be* some Great Cataclysm that purifies humanity prior to the realization of techno-utopia. However, it does come close. For example, Bostrom (2014) argues that the creation of machine superintelligence appears to constitute a "step risk," or a risk associated with (and only with) a great transformation of human history that is irreversible and, as mentioned above, will likely result in either total human annihilation or eternal paradise. Hence, it is not the case that the apocalypse is a survivable inflection point that we must cross to reach heaven on Earth, but that the creation of superintelligent machines could result in an apocalyptic event that terminates in the non-existence of all human beings. Utopia and annihilation are bound up together by an exclusive "or": if Utopia, then not-annihilation; but if not-Utopia, then annihilation. This could be sufficient, though, to foment apocalyptic fervor as advancements in computer science lead a greater proportion of Bostromians (or "longtermists") to anticipate the creation of superintelligence being imminent. If the research team closest to creating superintelligence is not perceived as taking seriously enough the safety concerns of existential risk scholars, a violent strike against this team may be, from a utilitarian perspective, entirely warranted. Consider Thomas Nagel's (1972) reflection on utilitarian thinking in the context of war: "Once the door is opened to calculations of utility and national interest, the usual speculations about the future of freedom, peace, and economic prosperity can be brought to bear to ease the consciences of those

responsible for a certain number of charred babies." In the case of superintelligence, it is not the "national interest" but the supposed "utopian interest" of humanity that is at stake. This could certainly excuse, ethically, more than "a certain number of charred babies." If an omelet you want, a few eggs must crack. Bostrom himself (2002) has argued that existential risk believers should "retain a last-resort readiness for preemptive action." In his words:

> Creating a broad-based consensus among the world's nation states [e.g., to "pass and enforce national laws against the creation of some specific destructive version of nanotechnology"] is time-consuming, difficult, and in many instances impossible. We must therefore recognize the possibility that cases may arise in which a powerful nation or a coalition of states needs to act unilaterally for its own and the common interest [i.e., reaching Utopia]. Such unilateral action may infringe on the sovereignty of other nations and may need to be done preemptively.

This way of viewing cosmic history, costs and benefits, becomes increasingly worrisome as the size and influence of the community grows.[9] This point is nicely articulated by one scholar as follows in discussing various small-scale utopian movements of the 1800s:

> Most of these 19th-century utopian experiments were relatively harmless because, without large numbers of members, they lacked political and economic power. But add those factors, and utopian dreamers can turn into dystopian murderers. People act on their beliefs, and if you believe that the only thing preventing you and/or your family, clan, tribe, race, or religion from … achieving heaven on Earth … is someone else or some other

> group, then actions know no bounds. From homicide to genocide, the murder of others in the name of some religious or ideological belief accounts for the high body counts in history's conflicts, from the Crusades, Inquisition, witch crazes, and religious wars of centuries gone to the religious cults, world wars, pogroms, and genocides of the past century (Shermer 2018[10]).

In fact, interest in existential risks has grown significantly in part due to the EA (Effective Altruism) movement, which has given rise to an ethical-ideological position called "longtermism." This is, according to Benjamin Todd (2017), equivalent to the "long-term value thesis," which states that what matters most, ethically, is not how things are going right now but how they will go in the long run. Although one need not be a utilitarian to be an EA, many EAs are utilitarians or at least are "most sympathetic to utilitarianism," as MacAskill observes (DS 2019). Indeed, the founder of Giving What We Can (GWWC) Toby Ord argues that utilitarianism, along with the Scientific Revolution and eighteenth-century Enlightenment, has "greatly contributed to the upbringing of effective altruism" (Ord and MacAskill 2016). And before the fledgling EA movement of "super-hardcore do-gooders" voted on the appellation "effective altruism," it seriously considered the appellation "effective utilitarian community" (EUC) (MacAskill 2014). The EA movement has raked in large quantities of money from wealthy "philanthropists," and consequently it has elevated the public (and academic) visibility of longtermism. Since longtermism is more or less equivalent with Bostromianism, it has thus played a crucial role in promulgating this utopian worldview.[11]

Finally, it is worth noting that many of the most violent apocalyptic movements in the past began as peaceable, irenic, passive groups. As Richard Landes (2011) notes,

> millennialism is a dynamic phenomenon, and in the course of an apocalyptic episode, a movement can literally flip from one extreme to the other. Among the classic cases, we find the Anabaptists who, in the course of their failed millennium at Münster from 1533-35, went from the most radically pacifist and egalitarian of the new "Protestant" groups to a violent and authoritarian group.

Another example is the Japanese doomsday cult Aum Shinrikyo. The group's leader, Shoko Asahara, initially predicted that Armageddon (or World War III—a kind of step-risk, so to speak), would occur in 1999. The initial aim of Aum was to *prepare the world* for this cosmic event by "teaching practices they believed would aid people in reducing negative karma for a better rebirth." But in the late 1980s, Asahara's views shifted from saving others to saving members of the group, who he believed "would be the sole survivors of a nuclear attack" that would destroy Japan, perpetrated by the US (Flannery 2016). By 1993 or 1994, he began to teach that Aum was not merely waiting for the coming grand battle but had a special role in *bringing Armageddon about*, which it later attempted to do with the 1995 Tokyo subway sarin attack. This Gestalt shift thus occurred over only a few short years, triggered by external circumstances. Similarly, The Covenant, The Sword, and the Arm of the Lord (CSA) began as a fundamentalist community "desiring to live a simple, pure Christian life in preparation for the endtime," i.e., utopia (Flannery 2016). But CSA's founder, the white supremacist James Ellison, later introduced Christian Identity teachings that flipped the apocalyptic Gestalt to an active mode according to which individuals came to believe that "I, as a righteous person, can trigger the end of the Evil age through my actions, especially through eliminating Evil on earth" (Flannery 2016). The point is that the

dangerousness of an ideology depends on two factors: the *content* of the ideology—Is it utopian? Does it encourage means-ends reasoning?—and the *context* in which that ideology is embedded—Is there an imminent threat that utopia will not materialize? Are we in the midst of an "apocalyptic episode"? Some of the examples above, such as the one vividly depicted by Häggström, illustrate how circumstances could come about such that a techno-utopian ideology that currently appears benign could lead to mass atrocities. The combination of utopianism and utilitarianism is dangerous: it has been in the past and it will be in the future.

3.2 Eurocentrism and White Supremacy. But there are other ways that the Bostromian paradigm can be harmful. For example, if one accepts that limited resources should not be "frittered away" on global catastrophic risks that do not pose existential risks, then if one comes to believe that, for example, climate change does not constitute an existential risk, one will conclude that it ought not constitute priority one, two, three, or four (or five). In fact, some longtermists like John Halstead (2020) have argued that climate change does not pose an existential risk—an idea picked up by *Vox* (Piper 2019)—and Bostrom himself (2019) has mostly dismissed this hazard, suggesting that it does not constitute "a truly civilizational threat." One implication is that this supports oppressive systems of white supremacy. By this term, I mean actions or policies that reinforce "racial subordination and maintaining a normalized White privilege" (Rollock and Gillborn 2011). Or as Frances Lee Ansley (1997) writes, the term does not

> allude only to the self-conscious racism of white supremacist hate groups [but also] to a political, economic and cultural system in which whites overwhelmingly control power and material resources, conscious and unconscious ideas of white superiority and entitle-

> ment are widespread, and relations of white dominance and non-white subordination are daily reenacted across a broad array of institutions and social settings.[12]

Consider that the worst effects of climate change will be felt by the Global South: the poorest people in the world, in countries most affected by the long and sordid legacy of western colonialism, imperialism, political meddling, and so on, will suffer the most in a catastrophically devastating whirlwind of extreme weather events, megadroughts, desertification, biodiversity loss, food supply disruptions, social upheaval, economic collapse, and political instability. In contrast, the Global North, which is largely responsible for the rise in atmospheric CO2, will be in a far better position to adapt to these unprecedented harms. Hence, dismissing climate change because it does not constitute an obstacle for creating Utopia reinforces unjust racial dynamics, and thus supports white supremacy. The same goes for statements made by the leading EA longtermist Nick Beckstead in a dissertation that has been widely cited and praised by the longtermist community. In this manuscript, titled "On the Overwhelming Importance of Shaping the Far Future" (2013), Beckstead argues that since what matters more than anything else is how well things go in the long run, from the point of view of the universe, then since people in rich countries are better positioned to shape the far future, their lives matter more than the lives of people in poor countries. As he makes the point,

> 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.

This is overtly white-supremacist.[13,14] It advocates a Eurocentric "give to the rich" policy based on the idea that huge numbers of currently non-existent, possibly never-existent people could clutter our future light cone, thus maximizing the total amount of impersonal intrinsic value that we could create in the long run. Indeed, Beckstead (with two co-authors) wrote in an EA article the same year as his dissertation:

> 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 (Singer, Beckstead, and Wage 2013).

Similar rhetoric is found among other prominent longtermists, such as Todd. For example, he writes that "since the future is big, there could be far more people in the future than in the present generation. This means that if you want to help people in general, your *key concern* shouldn't be to help the present generation, but to ensure that the future goes well in the long-term" (Todd

2017; italics in original). If one takes this seriously, then one should not be mainly concerned with helping poor people in the Global South who are suffering (and will suffer immensely) because of climatic anomalies induced by pollution produced mostly by the Global North. In my view, this is an unacceptable conclusion, morally speaking. It is, once again, predicated (more or less) on the view that the death of a living person is equivalent to the non-birth of a non-existent person; since so many non-existent persons (or value-containers) could come to exist in the future than exist in the present, shaping the future matters more than ameliorating the present (except insofar as ameliorating the present is necessary for creating all these possible people).

4. Conclusion

This paper is not merely a critique. It is also a warning: utopian ideologies, *especially* when mixed with overtly utilitarian ethical outlooks, can be (and often have been) extremely dangerous. Even when the progenitors of such belief systems have advocated a peaceful transformation from the present world to the future utopian otherworld, others have taken a more belligerent approach. As Bostrom (2008/2020) writes: "*What is Guilt in Utopia?* Guilt is our knowledge that we could have created Utopia sooner." For reasons stated above, the Bostromian worldview is a tinderbox. But even if it were not, there are still reasons for agreeing with Pinker's assessment of the concept as a "useless category." First, it is based on a techno-utopian vision that few academics, policymakers, or members of the general public would find appealing (not to mention that, as Harari argues, transhumanism could hugely exacerbate disparities between the socioeconomic elite and the rest). Second, given the axiological proposition that currently non-existent, possibly never-existent people have the same moral worth (as the containers

of value) as actual and future people, it lumps together atrocious and benign scenarios (like S1 and S2) as constituting the worst-case outcomes for humanity. For these reasons, I urge scholars not to adopt the theoretical framework of Bostromianism, and hence to reject his conception of *existential risk*. There are alternative frameworks that are far more compelling, and far less dangerous, than the dominant paradigm within ERS today.

References

Ansley, F.L. 1997. White Supremacy (And What We Should Do About It). In R. Delgado & J. Stefancic, J. (eds) (1997) *Critical White Studies: Looking Behind the Mirror*. Philadelphia, PA: Temple University Press, pp. 592-595.

Beckstead, Nick. 2013. *On the Overwhelming Importance of Shaping the Far Future*. Dissertation, Rutgers University.

Bostrom, Nick. 1997. How Long Before Superintelligence? *International Journal of Future Studies*. 2.

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

Bostrom, Nick. 2003. Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*. 15(3): 308-314.

Bostrom, Nick. 2005. Transhumanist Values. *Review of Contemporary Philosophy*. 4.

Bostrom, Nick. 2008/2020. Letter from Utopia. https://www.nickbostrom.com/utopia.html.

Bostrom, Nick. 2009. The Future of Humanity. *Geopolitics, History, and International Relations*. 1(2): 42-78.

Bostrom, Nick. 2013. Existential Risk Prevention as Global Priority. *Global Policy*. 4(1): 15-31.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chirot, Daniel, and Clark McCauley. 2006. *Why Not Kill Them All? The Logic and Prevention of Mass Political Murder*. Princeton, NJ: Princeton University Press.

Cirkovic, Milan. 2002. Cosmological Forecast and Its Practical Significance. *Journal of Transhumanism*. 12. https://www.jetpress.org/volume12/CosmologicalForecast.htm.

Cirkovic, Milan. 2019. Space colonization remains the only long-term option for humanity: A reply to Torres. *Futures*. 105: 166-173.

Deudney, Daniel. forthcoming. *Dark Skies*.

DS. 2019. How To Do The Most Good: An Interview With Will MacAskill. https://dailystoic.-com/will-macaskill-interview/.

Flannery, Frances. 2016. *Understanding Apocalyptic Terrorism: Countering the Radical Mindset*. New York, NY: Routledge.

Häggström, Olle. 2016. *Here Be Dragons: Science, Technology and the Future of Humanity*. Oxford: Oxford University Press.

Halstead, John. 2020. Is Climate Change an Existential Risk? https://docs.google.com/document/d/1qmHh-cshTCMT8LX0Y5wSQm8FMBhaxhQ8OlOeRLkXIF0/edit#.

Khatchadourian, Raffi. 2015. The Doomsday Invention. *The New Yorker*. https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

Kuhlemann, Karin. 2019. Complexity, Creeping Normalcy, and Conceit: Sexy and Unsexy Catastrophic Risks. *Foresight*. 21(1): 35-52.

Kupferschmidt, Kai. 2018. Taiming the Monsters of Tomorrow. *Science*. https://www.sciencemag.org/news/2018/01/could-science-destroy-world-these-scholars-want-save-us-modern-day-frankenstein.

Kurzweil, Ray. 2005. *The Singularity is Near*. New York: Penguin Books.

Kurzweil, Ray. 2006. Reinventing Humanity: The Future of Human-Machine Intelligence. *The Futurist*. https://www.kurzweilai.net/reinventing-humanity-the-future-of-human-machine-intelligence.

Landes, Richard. 2011. *Heaven on Earth: The Varieties of the Millennial Experience*. Oxford: Oxford University Press.

MacAskill, Will. 2014. The History of the Term "Effective Altruism." EA Forum. https://forum.effectivealtruism.org/posts/9a7xMXoSiQs3EYPA2/the-history-of-the-term-effective-altruism.

Nagel, Thomas. 1972. War and Massacre. *Philosophy & Public Affairs*. 1(2): 123-144.

Ord, Toby, and Will MacAskill. Opening Keynote, EA Global 2016. https://www.youtube.com/watch?v=VH2LhSod1M4&t=1339s.

Pinker, Steven. 2011. *The Better Angels of Our Nature: Why Violence Has Declined*. New York, NY: Viking Books.

Piper, Kelsey. 2019. Is Climate Change an "Existential Threat"—Or Just a Catastrophic One? https://www.vox.com/future-perfect/2019/6/13/18660548/climate-change-human-civilization-existential-risk.

Rollock, Nicola, and David Gillborn. 2011. Critical Race Theory (CRT). British Educational Research Association online resource. http://www.bera.ac.uk/files/2011/10/Critical-Race-Theory.pdf.

Shermer, Michael. 2018. Utopia Is a Dangerous Ideal. We Should Aim for "Protopia." *Quartz*. https://qz.com/1243042/utopia-is-a-dangerous-ideal-we-should-aim-for-protopia/.

Singer, Peter. 2015. *The Most Good You Can Do: How Effective Altruism Is Changing Ideas About Living Ethically*. New Haven, CT: Yale University Press.

Singer, Peter, Nick Beckstead, and Matthew Wage. 2013. Preventing Human Extinction. EA Forum. https://forum.effectivealtruism.org/posts/tXoE6wrEQv7GoDivb/preventing-human-extinction.

Srinivasan, Amia. 2015. Stop the Robot Apocalypse. *London Review of Books*. https://www.lrb.co.uk/the-paper/v37/n18/amia-srinivasan/stop-the-robot-apocalypse.

Tomasik, Brian. 2017. How Likely is a Far-Future Utopia? https://reducing-suffering.org/utopia/.

Verdoux, Philippe. 2009. Transhumanism, Progress and the Future. *Journal of Evolution and Technology*. 20(2): 49-69.

[1] Note here that "going extinct" need not entail "not surviving." There are many possible ways for humanity to go extinct but still survive, a point that has not yet been recognized in the relevant literature.

[2] See Verdoux 2009.

[3] However, in 2008 he added a postscript to this, which states: "I should clarify what I meant when in the abstract I said I would "outline the case for believing that we will have superhuman artificial intelligence within the first third of the next [i.e. the this] century." I chose the word "case" deliberately: In particular, by outlinining [sic] "the case for," I did not mean to deny that one could also outline a case against. In fact, I would all-things-considered assign less than a 50% probability to superintelligence being developed by 2033. I do think there is great uncertainty about whether and when it might happen, and that one should take seriously the possibility that it might happen by then, because of the kinds of consideration outlined in this paper."

[4] One might object here that the epistemological status of Bostrom's futurology is fundamentally different from the epistemological status of religious eschatologies. But this would be wrong for the following reason: the risks facing humanity are indeed based on the best science and evidence, whereas prophecies of Armageddon, the apocalypse, the end of the world, the eschaton, and so on, found in religious texts are based on faith and revelation. That makes worrying about, for example, human extinction as a result of an asteroid impact entirely different from worrying about, for example, the Antichrist. But this is not the case with transhumanism: this is a utopian value system that, as such, is not based on empirical, scientific truth. It is not a belief but a desire. There is no fact of the matter, no possible process of intersubjective verification, that could arbitrate between transhumanists and non-transhumanists. This matters because if "existential risk" is defined in terms of utopia, and if utopia is characterized in terms of transhumanism, then "existential risk" is defined in terms of a non-scientific value system.

[5] For a nice discussion of the topic, see Tomasik 2017. I personally much prefer the concept of *protopia* to *utopia*, just as I prefer the Russellian notion of *universal death* (human extinction) to Bostrom's notion of *existential risk*.

[6] Singer (2015) refers to Bostrom's position as entailing "harsh language."

[7] See Kuhlemann 2019.

[8] Although John Stuart Mill embraced a form of "utopian socialism" (Montgomery 2011).

[9] Perhaps because of the "unilateralist's curse" (Bostrom et al. 2016).

[10] Please note that Michael Shermer, the author of this passage, has been accused by multiple women of sexual harassment, assault, and even rape.

[11] Billionaires like Peter Thiel, a libertarian (and neoreactionary) who believes that women's rights "have rendered the notion of 'capitalist democracy' into an oxymoron," has given talks at EA events, and prominent EAs like Matt Wage have taken lucrative jobs on Wall Street to give more money to EA-approved charities—an idea called "earn to give." Furthermore, a "top donor" to the Future of Life Institute, which has strong ties to Bostromian and longtermist groups, is Sam Harris, who (as will become relevant below) has argued that black people are less intelligent than white people because of genetic evolution. I am not here leveling an *ad hominem* attack against Bostromianism/longtermism because of its association with such people. To the contrary, Bostromianism/longtermism attracts such figures because it appeals to their concerns.

[12] Similarly, Bell Hooks (1997) writes: "To me an important breakthrough, I felt, in my work and that of others was the call to use the term white supremacy, over racism because racism in and of itself did not really allow for a discourse of colonization and decolonization, the recognition of the internalized racism within people of color and it was always in a sense keeping things at the level at which whiteness and white people remained at the center of the discussion. In my classroom I might say to students that you know that when we use the term white supremacy it doesn't just evoke white people, it evokes a political world that we can all frame ourselves in relationship to."

[13] The prevalence of such tendencies within EA longtermism (and the invisibility of these tendencies among members) may be somewhat unsurprising given that, as of 2017, 89 percent of EAs are White, 7 percent are Asian, 3 percent are Hispanic, and only 1 are Black. The community is also dominated by men (70 percent).

[14] Thanks to Azita Chellappoo for apprising me of these papers.