Were the Great Tragedies of History "Mere Ripples"? The Case Against Longtermism

Phil Torres

Chapter 1: Longtermism

If you're the type of person who follows public "intellectuals" like Sam Harris, browses popular media like *The New Yorker* and *Vox*, or hopes to do the most good in the world, you have very likely heard about *longtermism*.[1] It is one of the central ideas in Toby Ord's popular new book *The Precipice*, published in 2020, and is closely linked to the concept of an *existential risk*. Not only has the term become more visible to the public over the past few years—and longtermists have big plans for this trend to continue—but projects associated with longtermism have, over just the last year, received literally millions of dollars in funding. Anecdotally, I have noticed a rapidly growing number of young people and established scholars flocking to the new field of Existential Risk Studies, which is largely motivated by longtermist ideas.

In this mini-book, written for students, journalists, and academics curious about this new ideology, I want to explain why longtermism—at least in its most influential guises—could be extremely dangerous. As outlined in the scholarly literature, it has all the ideological ingredients needed to justify a genocidal catastrophe. If this sounds hyperbolic, then keep reading. I strongly suspect that by the end of what follows you'll come to agree, or at least acknowledge that this ideological package is a ticking time-bomb. Hence, this mini-book is not just a critique but a *warning*: longtermism is a radical ideology that could have disastrous consequences if the wrong people—powerful politicians or even lone actors—were to take its central claims seriously.

There are many different definitions of "longtermism," all of which have in common a pivot toward taking seriously the long-term future of humanity.[2] This by itself sounds very appealing, and I believe it *should* sound this way. The world faces many problems that cannot be solved without thinking hard about the future—not just to

the next quarterly report, or the next election, or the lifetime of one's grandchildren, but *centuries* henceforth. To overcome the "Great Challenges" facing our species, we need more foresight and forecasting, more sober reflection on the potential causes and moral implications of human extinction.

Longtermism, though, goes far beyond a simple shift away from the myopic, short-term thinking that plagues our contemporary milieu. In what follows, I will focus on a group of ideas that have greatly shaped contemporary longtermist ideologies. We can label this *Bostromism* after its progenitor, the Oxford philosopher Nick Bostrom. It is this vision of what humanity's future ought to be that I worry about. It is a vision that, as we will see, commands us to subjugate nature, maximize economic productivity, colonize space, build vast computer simulations, create astronomical numbers of artificial beings, and replace humanity with a superior race of radically "enhanced" posthumans. Its basic tenets imply that the worst atrocities in human history fade into moral nothingness when one takes the big-picture view of our cosmic "potential," that preemptive war can be acceptable, that mass invasive surveillance may be necessary to avoid omnicide, and that we should give to the rich instead of the poor. However bad worldwide poverty and factory farming may be, solving these realtime global catastrophes aren't in our top five global priorities. In a catch-22, Bostromism adds that *not* developing technology would constitute an existential catastrophe in itself, even though the *primary reason* we face an estimated 20-percent chance of extinction this century is "technological progress."

In many cases, these claims are *explicit* in the writings of Bostrom and other longtermists. No inference is necessary: they are right there in black and white. I know because until recently I've been an enthusiastic participant in the research community, even writing the first introductory textbook on existential risks. But the more I worked on the topic, the longer I spent reflecting on the un-

derlying assumptions, the clearer it became that the nucleus of Existential Risk Studies—the Bostromian version of longtermism—could justify a wide range of unthinkable crimes against humanity. The multiplicity of dangers connected to this way of thinking about morality and the future were further reinforced by previous research that I'd conducted on apocalypticism and religious eschatology (where "eschatology" roughly means "the study of the world's end"). The fact is that longtermism has strong *millennialist* tendencies. If history is our guide, this makes it vulnerable to flipping from a passive to an active, violent mode of bringing about the end of the world —meaning ushering in the techno-utopian world dreamed of above.

Indeed, the parallels between apocalyptic religion and longtermism are striking. For longtermists, we stand at the most pivotal moment in human history—"the Precipice," as Ord calls it—that will determine whether the future is filled with near-infinite amounts of goodness or an empty vacuum of unforgivable moral ruination. This century is the "Grand Battle" that must be won at all costs—a directive given to us not by God but by the utilitarian imperative to maximize value as seen from "the point of view of the universe." If we win this battle, then the probability of extinction will drop close to zero and the paradise described above will be within reach. If we lose it, then all will be lost.

Because I want to keep everything as short as possible, the following chapters will outline the bare minimum of what readers need to know about the two main gears in the Bostromian machine: transhumanism and "total" utilitarianism. I will then piece together how this position could pose profound dangers.

Chapter 2: Surpassing Bliss and Delight

According to Bostrom's paper "A History of Transhumanist Thought," first published in 2005, the term "transhumanism" was coined by Julian Huxley in his 1927 book *Religion without Revelation*. But this is not true. The first time that Huxley used the term was in a 1951 lecture, although it was employed in roughly its contemporary sense even earlier, in 1940, by a Canadian historian and philosopher. Nonetheless, Huxley clearly stated in his 1957 book *New Bottles for New Wine*, building upon ideas expounded in 1927, "the human species can, if it wishes, transcend itself—not just sporadically, an individual here in one way, an individual there in another way—but in its entirety, as humanity." Huxley was a leading eugenicist who identified forced sterilization, demographics, and scientific knowledge of the genetic basis of intelligence as the means by which humanity could "transcend itself," by which he meant "realizing … new possibilities of and for his human nature." Obviously, this way of thinking opened the door to the Nazi atrocities of the Second World War, in which Hitler's diabolical regime force sterilized more than 400,000 people.

After WWII, most people wanted nothing to do with eugenics. But in the waning decades of the twentieth century, this began to change. The catalyst was the exponential growth of genetic engineering technologies along with grand proclamations by futurists about the promises of nanotechnology and artificial intelligence (AI). This foregrounded a new method of human perfectibility: integrating biology and technology, organism and artifact, to create what two authors in 1960 called "cyborgs." Consequently, a new transhumanist movement began to coalesce in the late 1980s, enabled by the Internet and initially referring to themselves as *extropians*. They imagined completely reengineering the human being to yield one or more new superintelligent, immortal, ultra-wise, hyper-moral species of

posthuman beings—call them *Homo cyborgensis*, *Homo supersapiens*, or to borrow Yuval Noah Harari's term, *Homo deus*, meaning "human god." As Bostrom wrote in 2003,

> our own current mode of being … spans but a minute subspace of what is possible or permitted by the physical constraints of the universe … It is not farfetched to suppose that there are parts of this larger space that represent extremely valuable ways of living, relating, feeling, and thinking.

This triggered a flurry of techno-utopian visions of a posthuman future in which our progeny live lives overflowing with pleasure and ecstasy.[3] Bostrom offers a tantalizing glimpse of this magical future in his "Letter from Utopia," first circulated in 2006 and later updated in 2020. It is composed by a fictional posthuman to his human ancestors (us), and hence is addressed "Dear Human" and signed "Your Possible Future Self." The posthuman author opens with the rhetorical question: "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." An effusive ballet of phantasmagoric imagery follows:

> 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 mar-

> velous 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.

The inspired weaver of words continues:

> 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!

How can humanity make this marvelous future a reality? How can we build this techno-utopian playground awash "in the pulsing ecstasy of love"? The posthuman tells us: "To reach Utopia, you must discover the means to three fundamental transformations." The first is that we must become *immortal* through life-extension technologies, which could include biomedical interventions of our bodies or uploading our minds to a computer. The second is that we must become *superintelligent*, since "it is in the spacetime of awareness that Utopia will exist." And the third is that we must elevate *well-being*, which a "hedonist" would equate with *pleasure*. "A few grains of this magic ingredient," the posthuman writes, "are worth more than a king's treasure."

Bostrom wasn't the only transhumanist with intoxicating hopes for a better world to come—a heavenly otherworld built by science and technology rather than supernatural forces. The other most prominent transhumanist so far this century is Ray Kurzweil, author of *The Singularity is Near* (2005). Whereas the early transhumanists

called themselves "extropians," Kurzweil espoused a version called "singularitarianism." This emphasized the historical *discontinuity* that creating advanced AI systems would bring about. Bostrom himself has frequently suggested that the creation of superintelligent machines will result in either a utopian paradise or total human annihilation—a drastic situation reminiscent of the cosmic battles described in religious texts where the stakes are all-or-nothing. But Kurzweil made the stronger claim that in exactly 2045 "human life will be irreversibly transformed" as human and machine intelligence merge, resulting in non-biological forms of intelligence dominating the universe. At the same time, the rate of technological progress will accelerate beyond human comprehension. This is called the "technological Singularity"—or, derogatorily, the *techno-rapture*—and it will make it possible for 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.

Many other transhumanists put forward their own wide-eyed prognostications of the promises of tomorrow. For example, David Pearce—who co-founded the World Transhumanist Association with Bostrom in 1998—argues that we should reengineer not just the human organism but *all sentient beings in the biosphere*. He calls this the Abolitionist Project.

Along similar lines, the AI theorist Ben Goertzel defends an ideology called *Cosmism*, which has roots in the work of Nikolai Fy-

odorovich Fyodorov, a nineteenth-century Russian philosopher who advocated for radically extending our lives, resurrecting the dead, and other marvels. On Goertzel's view, *Cosmism* affirms the desirability of pursuing super-powerful advanced technologies to enable humans to fully 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, Goertzel writes, "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.'"

This is the first component of Bostromism. It offers one reason that failing to develop powerful new emerging technologies would constitute a "disaster." As Bostrom writes in his "Transhumanist Values" paper, the "core value" of transhumanism is "having the opportunity to explore the transhuman and posthuman realms." Exploring such realms is the only way that humanity can attain the techno-utopian world so delicately depicted by the posthuman author of Bostrom's "Letter from Utopia." Yet to realize this core value, we must develop what the transhumanist Mark Walker calls "person-engineering" technologies associated with genetic engineering, nanotechnology, and AI. Hence, the only way forward is more technology, despite the unprecedented hazards that they will introduce.

Chapter 3: The Mandate to Maximize

On the website utilitarianism.net, Will MacAskill and Darius Meissner write that "advocates of utilitarianism have argued that the theory has attractive theoretical virtues such as simplicity." But I find this misleading: utilitarianism is actually a complex bundle of many different ideas.[4] In this short chapter, I will do my best to outline the features and properties of utilitarianism that are most relevant to the critique below.

The first thing to note is that the question "What is right and wrong?" is *different* from the question "What is good and bad?" You may have heard the old saying that "whereas deontology puts the right before the good, consequentialism puts the good before the right." What this means is that for someone like Immanuel Kant, a deontologist *par excellence*, whether the consequences of an act are good or bad has nothing to do with whether the act is right or wrong. This is why he argued that it would be immoral to lie to a murderer at your door asking for the whereabouts of his next victim. Sure, the victim being murdered would be bad—Kant would agree—but what does that have to do with morality? So long as you follow unbendable moral rules like "Never lie," you're doing what's necessary to keep your moral house in order.

Utilitarians disagree. For them, you can't answer the first question above without already having answered the second one. The reason is that morally right actions are defined as *precisely those* that increase the amount of good in the world. But what is "the good"? There are several possible answers, but for the sake of our discussion, let's accept a hedonistic theory according to which the one and only intrinsically good thing in the universe is that subjective state called *pleasure* or *happiness*. This leads to the following definition: an act is morally right if and only if it maximizes the total amount of pleasure in the universe (compared to all the other acts available at the time).

This is the heart and soul of "total" utilitarianism, where "total" refers to the fact that what matters is the *absolute* quantity of pleasure rather than the *average*.[5] So, whereas Kant would claim that lying is wrong because it violates the Moral Law, utilitarians would say that lying is wrong when and only when doing so would fail to maximize intrinsic value—that is, pleasure, if one's a hedonist.

But how should we assess whether an act has in fact maximized intrinsic value? Consider this somewhat silly example: you hack into the bank accounts of 1,000 people and steal $100 from each. You then take the resulting $100,000 and spend it on a relaxing vacation on the sunny tropical beaches of a Caribbean island. Did stealing in this case maximize the good? Clearly, the total amount of pleasure has increased *for you*. But ethics is supposed to be impartial and objective, not biased toward particular individuals. It strives to assess our moral choices and actions from a neutral perspective called the *moral point of view*.[6] Ethicists have proposed many different accounts of what the moral point of view should be. For Kant, it was the Categorical Imperative. But for Henry Sidgwick, an influential early classical utilitarian, the moral point of view is nothing more or less than "the point of view of the universe." Accordingly, we should assess the consequences of actions from a disembodied cosmic eye, which, looking down from above, can calculate the overall increase or decrease in pleasure objectively. In the case above, although sneaking away with $100,000 was good overall for you, it was not good overall *for the universe*, so to speak, which considers the effects of an action on everyone equally.

So far so good! There is one more aspect of this view, especially as Bostrom understands it, that we need to establish. One of the many criticisms of utilitarianism is that it's insensitive to the distinction between persons. Consider a situation in which you're told that if you get a slightly painful medical procedure next week, you can save yourself from an extremely painful procedure in 1 year. Many of

us would opt for the procedure next week. In doing this, we inflict pain upon ourselves so that our future self can avoid it. For utilitarians, though, we should think about trade-offs *between* lives in the exact same way: if inflicting pain upon one person next week would enable a *separate* person from experiencing far worse pain in 1 year, then morality orders us to do it. This is what "the greater good" as seen from the cosmic vantage point is all about: making trade-offs here and there without thinking about persons as inviolable beings, or actions as being constrained by the sorts of rules that Kant proposed.[7]

On this view, persons—you and me, your grandma and your spouse—do not matter *in and of ourselves*.[8] We are *mere means to an end*, that end being maximal pleasure in the world. As John Rawls famously put it, persons are just the "containers" of intrinsic value. We are fungible (that is, interchangeable) receptacles that matter only because pleasure cannot exist without containers to contain it.[9] This leads to a startling conclusion: the death of someone you love dearly is no worse, morally speaking, than the non-birth of someone who could have existed but never will. To illustrate, think of a person in your life who you love dearly, and imagine that person perishing. (Sorry for the dark thought!) Now imagine a merely possible person named "Diego." He is what I would call a currently non-existent, possibly never-existent imaginary being. Let's say that, in fact, he never comes into existence. Which of these two scenarios is worse? Bostrom would say that they're equivalent, given that (a) Diego would have a happy life, and (b) we bracket the *extra* suffering that your loved-one's death would cause those who survive. In other words, the *death itself* is morally equivalent to the *non-birth* of Diego. Why? The answer should be obvious: your loved one and Diego are just containers, and in terms of the total amount of pleasure in the universe, there's no fundamental difference between removing a container that exists and failing to create a container that could exist.

To tie these threads together, morality abhors a vacuum. The more value containers that exist, the more potential value. The more value, the better the world becomes. Hence, this utilitarian view commands us to maximize the total number of containers—meaning "people"—with net-positive amounts of value. The best possible outcome is one in which the largest number of happy people exist. Bigger equals better.[10]

Chapter 4: Astronomical Value and Existential Risks

The question then is: How many future people could there be? In short, a lot. The first to crunch the numbers was Carl Sagan in a 1983 article published in *Foreign Affairs*. He calculated that if humanity remains on Earth and survives "over a typical time period for the biological evolution of a successful species," which he specified as 10 million years, and if the human population remains stable at 4.6 billion (the number of people in 1983), then some *500 trillion people* may yet come into existence. This is why he argued that "the stakes are one million times greater for extinction than for the more modest nuclear wars that kill 'only' hundreds of millions of people."

But why would we remain on Earth? If what matters is maximizing value-containers, why not spread into our *future light cone*, or the region of spacetime that is theoretically accessible to us at any given moment traveling at the speed of light. In a 2003 paper in the *Journal of Transhumanism*, which seems to have drawn from an earlier paper by Milan Ćirković, Bostrom concluded that about 10^{23} biological humans could come to exist within the Virgo Supercluster alone.[11] The Virgo Supercluster is a giant cosmic structure that contains about 100 galaxy groups, one of which is our own Local Group, which includes at least 80 distinct galaxies, one of which is our own Milky Way. Yet there are some *10 million* superclusters in the observable universe, and while not all of these may be reachable to us given the expansion of the universe, the mathematical implications are clear: the future population of intergalactically spacefaring posthumans could be ginormous.

But why would we remain biological? If simulated beings can have conscious experiences of pleasure, then they can be containers no less than us. So, imagine this: our descendants fly out into the cosmos and convert every exoplanet they encounter into *computronium*, which refers to a configuration of matter and energy that is op-

timized to perform computational tasks like—drum roll—simulating conscious minds. These descendants then design high-resolution simulation worlds in which they plop massive numbers of simulated beings living, as Bostrom puts it, "rich and happy lives while interacting with one another in virtual environments." (Note that Bostrom never tells us *why* these people, perhaps knowing full well that they're living in simulated worlds, are happy. Maybe they're utilitarians who understand that it's their moral duty to be happy for the sake of adding more intrinsic value to the universe. Or maybe there is some sort of digital Prozac that they can get from their local digital pharmacy.) If this were to happen, Bostrom joyfully reports that some 10^{58} conscious beings—that's a 1 followed by 58 zeros!—with lifespans of 100 years could exist thanks to these simulations, although "the true number is probably larger." The point, as he noted in 2003, is "not the exact numbers but the fact that they are huge."

What does all of this mean? It means that the total amount of intrinsic value that could come to exist within our future light cone could be *astronomically large*. I call this the "astronomical value thesis." It further implies that, since morality is built upon value, according to utilitarianism, we have an overriding, profound moral obligation to ensure that as many of these currently non-existent, possibly never-existent people *are actually born*.

The next question is *practical*: how exactly could we accomplish this? We have already mentioned that one important step is colonizing space. Without doing this, the total human or posthuman population will be severely limited by the carrying capacity and resources of our tiny planetary oasis. But is there more?

Bostrom answers this question in his 2013 paper titled "Existential Risk Prevention as Global Priority." (Note that the paper's title on Bostrom's website is different.) To *maximally maximize* intrinsic value, we must reach and sustain what he calls "technological maturity," which denotes "the attainment of capabilities affording a level of

economic productivity and control over nature close to the maximum that could feasibly be achieved." Once we have increased economic productivity and subjugated the natural world to the physical limits (insofar as this is feasible), we will be able to maximally harness all of the universe's vast resources—our so-called "cosmic endowment" of negentropy—which await our eager plundering. With all of this free energy in hand, with every star and galaxy and supercluster subdued within the kingdom of posthuman hegemony, the grand desiderata of transhumanism and utilitarianism can be fulfilled. That is, technological maturity would allow us to explore every corner of the posthuman realm (the core value of transhumanism) and run the maximum number of simulations full of trillions and trillions (and trillions and trillions) of conscious beings.[12]

This leads Bostrom to define "existential risk" in terms of technological maturity. In essence, this encompasses *any* future event that would either permanently prevent us from reaching technological maturity *or* cause us to lose technological maturity after achieving it. The most obvious way that this could happen is for humanity to go extinct. But there are a plethora of *survivable* scenarios as well. Bostrom thus proposes a four-part classification of existential risk "failure modes," which goes as follows (to quote him):

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

So, to sum up: transhumanism outlines a picture of what Utopia would look like for *individuals*. It is a place in which posthuman beings are bestowed with superintelligent minds, total control over their emotions, indefinitely long lifespans, experiences saturated with ecstasy, and other superhuman delights. Utilitarianism offers an account of "utopia" from *the point of view of the universe*. It is a configuration in which the cosmos is overflowing with intrinsic value, value, value, value—impersonally conceived. To realize these overlapping utopias, we must attain a stable state of technological maturity, and failing to do this would constitute an existential catastrophe—the worst possible outcome for not just humanity but the Sidgwickian universe itself.

Let's now turn to some of the implications of this Bostromian view.

Chapter 5: Bostrom's Altruist

We begin with the following argument:

(p1) Since an existential catastrophe would prevent astronomical numbers of people from coming into existence, and
(p2) since we have an overriding, profound obligation to ensure that astronomical numbers of people come to exist, it follows that
(c) we have an overriding, profound obligation to avoid an existential catastrophe.

As Bostrom wrote in 2013, the potential size of the future implies "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." He formalizes this idea in a "rule of thumb for such impersonal moral action" that he calls the *maxipok rule*. This states that we must "maximize the probability of an 'OK outcome,' where an OK outcome is any outcome that avoids existential catastrophe." Earlier, in 2003, he made the point in writing that "for standard utilitarians, priority number one, two, three and four should … be to reduce existential risk." This was stated in the paper that provided his initial calculations of how many people could exist in the future if we were to colonize the Virgo supercluster and create simulated universes. Hence, one is left to surmise that the fifth priority should be to venture beyond Earth and begin simulating conscious beings. The point is driven home by Bostrom's calculation that "approximately 10^{38} lives [are] lost every century that colonization of our local supercluster is delayed," which equals "about 10^{29} potential human lives per second." In other words,

every single second that we fail to spread through and exploit the material resources of the Virgo Supercluster entails what he called "astronomical waste," i.e., huge quantities (in absolute terms) of our cosmic endowment of negentropy dissipating forever.

What isn't among the top five priorities for humanity? Well, everything that *doesn't* reduce the probability of an existential catastrophe or get ourselves ready to colonize the universe. This includes alleviating global poverty, eliminating factory farming, and even mitigating climate change, since climate change is not seen by (most) Bostromians as an existential catastrophe. (It might be a setback, but it probably won't instantiate any of the failure modes above.) As Bostrom writes,

> 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, we would do well to focus on this most efficient philanthropy.

In a commentary on this passage, Peter Singer describes Bostrom's dismissal of "donating to help the global poor or reduce animal suffering as a 'feel-good project' on which resources are 'frittered away'" as "harsh language" that is "likely to be counterproductive" to the longtermist cause. But of course Bostrom's contention makes perfect sense given the premises specified above. Indeed, Bostrom explicitly argues that the very worst disasters in human history fade to almost complete insignificance when one takes the cosmic view of our "potential" in the universe (i.e., subjugating nature and maximizing economic productivity). Referring to events like "Chernobyl, Bhopal, volcano eruptions, earthquakes, draughts [*sic*], World War I, World

War II, epidemics of influenza, smallpox, black plague, and AIDS," Bostrom writes that

> 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.

Mere ripples! That's what World War II—including the forced sterilizations mentioned above, the Holocaust that killed 6 million Jews, and the death of some 40 million civilians—is on the Bostromian view. This may sound extremely callous, but there are far more egregious claims of the sort. For example, Bostrom argues that the *tiniest reductions* in existential risk are morally equivalent to the lives of billions and billions of *actual human beings*. To illustrate the idea, consider the following forced-choice scenario:

> *Bostrom's altruist*: Imagine that you're sitting in front of two red buttons. If you push the first button, 1 billion living, breathing, actual people will *not* be electrocuted to death. If you push the second button, you will reduce the probability of an existential catastrophe by a teeny-tiny, barely noticeable, almost negligible amount. Which button should you push?

For Bostrom, the answer is absolutely obvious: you should push the second button! The issue isn't even close to debatable. As Bostrom writes in 2013, even if there is "a mere 1 per cent chance" that 10^{54} conscious beings living in computer simulations come to exist in the future, 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." So, take a billion human lives, multiply it by 100 billion, and what you get is the moral equivalent of reducing existential risk on the assumption that there is a "one billionth of one billionth of one percentage point" that we run vast simulations in which 10^{54} happy people reside. This means that, on Bostrom's view, you would be a *grotesque moral monster* not to push the second button. Sacrifice those people! Think of all the value that would be lost if you don't!

Many longtermists have come to agree with this reasoning. For example, Hilary Greaves, a utilitarian philosopher and colleague of Bostrom's at Oxford, writes that "a change that … increases the chance of imminent extinction by 0.00001%, is roughly welfare-equivalent to the intrinsic badness of an event that wipes out 10% of the population throughout the next century." In other words, 10 percent of the population dying is no different than the probability of human extinction rising by 0.00001 percent. Why? Because what matters is the total amount of intrinsic value within our future light cone, and without people—the containers of value—only a fraction of this impersonal value will ever exist.

Philosophers have a funny saying that goes: "One person's *modus ponens* is another persons *modus tollens*." Modus ponens and modus tollens are two rules of inference in propositional logic. Basically, someone might say, "I accept this conclusion *because* I accept the premises," whereas someone else might say, "I reject at least one of these premises *because* I reject the conclusion."[13] In the cases above, you should reject the premises because the *conclusions* are

patently absurd and wrong: the worst atrocities are not "mere ripples." The lives of billions of actual human beings are not morally equivalent to raising the probability of running exoplanetary simulations crowded with trillions of conscious beings in the future by minuscule amounts.

There is nothing wrong with moral theories that challenge our intuitions. Indeed, our moral intuitions have been wrong many times throughout history. Just consider that "cat burning" was once a popular form of entertainment in Europe. But reducing morality to economics, to mere number-crunching, involving fungible value containers, can have profoundly harmful real-world consequences. For example, consider the following scenario outlined by Olle Häggström in his 2016 book *Here Be Dragons*. He begins by asking us to 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 [or more] people." Häggström writes that

> I feel extremely uneasy about the prospect that [this line of reasoning] 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.

He continues:

> 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.

Perhaps a Bostromian longtermist could sidestep this unsavory conclusion by noting 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." Yet we should wonder "whether we can trust that our world leaders understand [such] points." Ultimately, Häggström makes the wise decision to abandon utilitarianism in favor of what ethicists would call an "absolutist" deontological constraint—that is, a non-negotiable restriction on "maximizing the good"—according to which "there are things *that you simply cannot do*, no matter how much future value is at stake!" (His italics.)

Unfortunately, not everyone would follow Häggström's sagacious lead. Not everyone would agree that violently murdering a large group of innocent people is *never* okay. Certainly not Bostrom's altruist. One could, indeed, easily justify the annihilation of Germany by paraphrasing Bostrom's own words: "Well, as tragic as this event would be to the Germans 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. Think about the situation from the universe's point of view! Imagine all the lost value if we fail to attain technological maturity!"

Chapter 6: Guilt in Utopia

The dangers of utilitarian modes of moral reasoning and the utopian promise of eternal life in paradise are well known. As Pinker writes,

> 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.

Although for Bostrom and other longtermists, the moral value is not *infinite*, it is *astronomical*. But this is more than enough, as we have seen, to justify pushing a button that trades billions of human lives for the tiniest reductions in existential risk. To quote the philosopher Thomas Nagel, writing about utilitarianism 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 longtermism, the issue is not the *mere* national interest of a state but the *Ultimate Moral Interest of the Universe itself*. How many charred babies is too many when the stakes are this high?

One might object at this point to the association of genocide, charred babies, and the like with longtermism. "My god, no one in

the longtermist community is calling for anything like this! It is simply inconceivable that anyone would!" There are several problems with this reply. First, even if the leading longtermists today would ultimately side with Häggström in rejecting extreme, violent measures to secure our future among the stars, so long as the ideology remains a "legitimate" option on the marketplace of ideas,[14] the door is open for others—long-term extremists who interpret Bostrom's words literally—to pick it up and run. In this sense, longtermism is an information hazard.[15]

But second, history is full of millennialist movements that rapidly flipped from one mode of eschatological thinking to another. A movement is *millennialist* if it holds that our current world is replete with suffering and death but will soon "be transformed into a perfect world of justice, peace, abundance, and mutual love." In many cases, members of such movements could not, initially, have imagined their community turning to violence. It would have been unthinkable. But when external conditions change in the right way, when new circumstances trigger something in the old ideology, a community's tolerance for bloodshed can quickly grow. As the historian Richard Landes observes,

> 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 [in Germany] from 1533-35, went from the most radically pacifist and egalitarian of the new "Protestant" groups to a violent and authoritarian group.

The Anabaptists are just one of an interminable number of examples from the dusty annals of history. For example, many readers may

find it surprising that the notorious Japanese doomsday cult Aum Shinrikyo actually began as a peaceful movement. Over the course of just a few years, it morphed into an active apocalyptic group that tried to initiate Armageddon by releasing the poisonous gas sarin into the Tokyo subway system in 1995. Since paradise awaits on the other side of the apocalypse, the sooner the apocalypse comes, the sooner believers will enter heaven.

Longtermism is, of course, a "secular" rather than "religious" movement, although one could very plausibly describe it as a *quasi*-religion whose central object of worship is not "God" but *future value*. Nonetheless, plenty of millennialist movements in recent history have been "secular," most notably those driven by the communist ideology of Marxism-Leninism. What matters is not whether the ideology is secular or religious, but whether it contains the ingredients necessary for an "apocalyptic episode" to instigate a violent response. In this case, an apocalyptic episode might be the scenario described by Häggström: putative knowledge that there is an omnicidal actor somewhere in Germany who's chances of success are 1 in a million. It is not difficult to imagine similar situations. All one needs is a moral commitment to Utopia and the belief that there is a greater-than-negligible chance that some actor could prevent us from reaching this prized future state. We do, after all, live in the most important moment of history—the Precipice—according to longtermists, and as the posthuman author of Bostrom's Letter from Utopia asks: "*What is Guilt in Utopia?* Guilt is our knowledge that we could have created Utopia sooner."

While Häggström's example is unrealistic right now—the only actors capable of unilaterally destroying the world are nation-states with large nuclear arsenals—emerging technologies will soon enable small groups and even lone wolves to wreak civilizational havoc.[16] Not only will this increase the number of people who could bring about an existential catastrophe, but it will increase the number of

people who could do things like obliterate Germany in an attempt to *prevent* an existential catastrophe. This means that the more prevalent longtermism becomes, the greater the chance of something like this happening.[17] Imagine a future in which groups all over the world initiate major attacks on other groups because they believe they might be plotting an omnicidal strike, and crunching the numbers reveals that the expected value of killing off any potential agent of omnicide is sufficiently large.

Of course, the possibility that emerging technologies could enable groups and individuals to destroy the world is *itself* a horrifying prospect—not because of the imaginary people who would never exist but because of the profound harm it could cause those living at the time. This is an issue that we'll return to below.

Chapter 7: Should the Rich Get Richer?

There are other reasons for worrying about longtermism gaining more clout. For example, consider Tyler Cowen's observation that utilitarianism seems to "support the transfer of resources from the poor to the rich … if we have a deep concern for the distant future." The reason pertains to the features of this ethical theory that we discussed above. The Oxford philosopher Andreas Mogensen echoes this idea in a more recent paper published by the Global Priorities Institute. "It has been assumed," Mogensen writes, "that utilitarianism concretely directs us to maximize welfare within a generation by transferring resources to people currently living in extreme poverty. In fact, utilitarianism seems to imply that any obligation to help people who are currently badly off is trumped by obligations to undertake actions targeted at improving the value of the long-term future."

Similarly, the longtermist Nick Beckstead wrote in his PhD dissertation from 2013 the following:

> 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.

In other words, since what matters most is shaping the far future—i.e., ensuring that the maximum amount of intrinsic value floods our future light cone to the brim—and since people in rich countries are in a better position to shape the far future than people in poor countries, the *lives* of people in rich countries matter more than the *lives* of people in poor countries. Longtermists like Ord have lavished praise on this dissertation, and it is cited on numerous webpages hosted by the Centre for Effective Altruism, which shares office space with the Future of Humanity Institute, founded by Bostrom in 2005.[18]

It is my hope that many readers will find the above statements problematic. This should be even more the case when one remembers that these arguments are based on the dubious contention that imaginary people—such as those Bostrom envisages living "happy lives" in computer simulations—are what tip the scale away from actual, living, breathing poor people in the world today. If what matters is the total amount of pleasure across space and time from the universe's disembodied perspective, then the fact that some 15,000 children die *each day* from hunger-related illnesses pales in comparison to the astronomical quantities of value that would be lost if an existential catastrophe were to occur. Children dying from starvation is sad, but trillions and trillions (and trillions and trillions) of value containers never coming into existence is *much*, *much sadder*.

Not only are these conclusions classist, but if implemented today they would support the ongoing dominance of the Global North in a world still recovering from the devastating effects of Western colonialism, imperialism, political meddling, exploitation, and so on. In a phrase, they support *white supremacist* ideology. To be clear, I am using this term in a technical scholarly sense. It denotes actions or policies that reinforce "racial subordination and maintaining a normalized White privilege." As the legal scholar Frances Lee Ansley

wrote in 1997, the concept encompasses "a political, economic and cultural system in which whites overwhelmingly control power and material resources," in which "conscious and unconscious ideas of white superiority and entitlement are widespread, and relations of white dominance and non-white subordination are daily reenacted across a broad array of institutions and social settings."

On this definition, the claims of Mogensen and Beckstead are clearly white supremacist: African nations, for example, are poorer than Sweden, so according to the reasoning above we should transfer resources from the former to the latter. You can fill in the blanks. Furthermore, since these claims derive from the central tenets of Bostromian longtermism itself, the very same accusation applies to longtermism as well. Once again, our top four global priorities, according to Bostrom, must be to reduce existential risk, with the fifth being to minimize "astronomical waste" by colonizing space as soon as possible. Since poor people are the *least* well-positioned to achieve these aims, it makes perfect sense that longtermists *should* ignore them. Hence, the more longtermists there are, the worse we might expect the plight of the poor to become.

Chapter 8: The Wrong Reflection?

Longtermism is one of the three main cause areas of the Effective Altruism (EA) movement.[19] Oddly enough, other major cause areas are alleviating global poverty and eliminating factory farming. Thus, there is a direct tension between longtermism, on the one hand, and these other two cause areas, on the other. In some cases, the tension is resolved by explicitly saying, as Beckstead does in his dissertation, that saving rich lives is "substantially more important" than saving poor lives, for the sake of the greater good over the extremely long term.

Others appear to be more tentative in their endorsement of longtermism. We should take the notion of *normative uncertainty* seriously, they claim, or the possibility that there are fundamental errors in our normative, including moral, beliefs. Yet leading longtermists like Bostrom and Ord are clear that at least some normative beliefs are non-negotiable. They are not up for debate. One example is "technological progress." This is so central to the longtermist vision that Bostrom identifies the *cessation* of further "progress" as an existential catastrophe that would instantiate the "permanent stagnation" failure mode listed above. Ord strongly agrees: "I don't for a moment think we should cease technological progress," he writes. "Indeed if some well-meaning regime locked in a permanent freeze on technology, that would probably itself be an existential catastrophe, preventing humanity from ever fulfilling its potential."

There are two main reasons that longtermists hold this view. The first concerns another non-negotiable commitment: humanity must do everything in its power to reach its "potential." For Bostrom, this means subjugating the natural world, maximizing economic productivity, simulating trillions (etc.) of conscious beings, and so on. For Ord, what constitutes our "potential" should be decided during a period that he calls the Long Reflection, which he imagines

commencing after the more immediate task of establishing Existential Security. I find these ideas so implausible that I won't here discuss them.[20] The point is that striving to fulfill our "potential" is not debatable, nor is the assertion that technology is the vehicle that will deliver us to this destination. As Ord puts it, "the best futures open to us—those that would truly fulfill our potential—will require technologies we haven't yet reached."

The second concerns the fact that we live in a hazardous universe—a veritable haunted house cluttered with death traps both above our heads and below our feet. This leads Ord to conclude that "without further technological progress we would eventually succumb to the background of natural risks such as asteroids." Technology, then, is necessary to avoid the otherwise inevitable extinction of humanity due to natural threats like asteroid and comet impacts, supervolcanic eruptions, gamma-ray bursts, galactic center outbursts, and so on. Yet at the same time, everyone agrees that by far the greatest source of danger to our collective survival is *technology itself*. Bostrom makes the point like this:

> The great bulk of existential risk in the foreseeable future is anthropogenic; that is, arising from human activity. In particular, most of the biggest existential risks seem to be linked to potential future technological breakthroughs that may radically expand our ability to manipulate the external world or our own biology. As our powers expand, so will the scale of their potential consequences—intended and unintended, positive and negative.

Indeed, many scholars within Existential Risk Studies agree that the probability of human extinction or civilizational collapse this century is significant. For example, Bostrom writes that his "subjective opinion is that setting this probability [of an existential catastrophe]

lower than 25% would be misguided, and the best estimate may be considerably higher." Later, during a TED talk, he claimed that "assigning a less than 20 percent probability would be a mistake in light of the current evidence we have."[21] In 2008, an informal survey of experts conducted by the Future of Humanity Institute put the median estimate of annihilation before 2100 at 19 percent. And in a 2017 interview, Ord says the because of "radical new technology," humanity has a mere 1 in 6 chance of surviving this century. Others, like Lord Martin Rees, believe that these new technologies give civilization a mere 50-50 chance of making it to the twenty-second century. A coin flip! Pause for a moment to allow these numbers to percolate between your wriggling neurons.

Now compare these estimates of catastrophe to the likelihood of extinction caused by natural threats. The most probable threat comes from supervolcanoes, which erupt on average once every 50,000 years. A supervolcano can spew sulfate aerosols into the stratosphere, which then spread around the world and block incoming solar radiation. This results in a decrease of photosynthesis and possible collapse of the food chains, leading to species extinctions. Yet humanity has survived *two* supervolcanic eruptions over the past 200-300 thousand years, which is our species' lifetime so far. These are the Toba eruption 75,000 years ago and the Oruanui eruption 26,500 years ago. What about an asteroid or comet impact, the other most probable threat? According to Bostrom, "this particular risk turns out to be very small. An impacting object would have to be considerably larger than 1 km in diameter to pose an existential risk. Fortunately, such objects hit the Earth less than once in 500,000 years on average."

So, think about the situation: without technology, we are vulnerable to (a) supervolcanoes exploding every *five hundred centuries* on average, two of which we have already survived with only stones and fire. And (b) impactors that strike Earth every *five thousand cen-*

turies on average. In contrast, because of technology, the probability of total human annihilation according to *longtermists themselves* hovers between 16.6 and 20 percent *this century*.

Arguing that we need more technology is just nuts. *The more technological we have become, the closer to self-annihilation we've inched*. In Bostrom's words, "with the exception of a species-destroying comet or asteroid impact (an extremely rare occurrence), there were probably no significant existential risks in human history until the mid-twentieth century." The implication here could not be more obvious—and there is no reason to believe the trend will reverse in the future.[22] And yet longtermists like Ord and Bostrom dig their heels in and dogmatically assert that *more technology is the answer*—that "we should not *blame* civilization or technology for imposing big existential risks," even though civilization and technology *are* responsible for the extremely dire predicament in which we find ourselves. Imagine boarding a plane and being told that it has a 20 percent chance of crashing. Would you get off? Sorry, let me rephrase that: would you get off by running or sprinting? This is humanity's situation right now—except that we are *already* 35,000 feet in the air, thanks to the triumphant strides of "technological progress" over the past seven decades.[23]

The craziness of this plight is not lost on all technophilic transhumanists. For example, Kurzweil writes the following in *The Singularity is Near*: "Imagine describing the dangers (atomic and hydrogen bombs for one thing) that exist today to people who lived a couple of hundred years ago. They would think it mad to take such risks." Yet Kurzweil tries to undermine this point with a fallacious argument that others, including Bostrom and Ord, make as well. "How many people," Kurzweil writes, "in 2005 would really want to go back to the short, brutish, disease-filled, poverty-stricken, disaster-prone lives that 99 percent of the human race struggled through a couple of centuries ago?" Similarly, Ord contends that

> technological progress has been one of the main sources of our modern prosperity and longevity—one of the main reasons extreme poverty has become the exception rather than the rule, and life expectancy has doubled since the Industrial Revolution. Indeed, we can see that over the centuries all the risks technology imposes on humans have been outweighed by the benefits it has brought.

First of all, the point of reference should not be "a couple of centuries ago" or "since the Industrial Revolution." As the anthropologist Mark Cohen writes in *Health and the Rise of Civilization*, "a good case can be made that urban European populations of that period may have been among the nutritionally most impoverished, the most disease-ridden, and the shortest-lived populations in human history." In fact, the Neolithic Revolution resulted in a significant decline in human health, as evidenced by a drop in the average height of populations; it was not until the mid-twentieth century that populations in the affluent West regained their lost verticality. Jared Diamond may not be off the mark when he describes the invention of farming as "the worst mistake in human history."

Second, to say that extreme poverty is now the exception depends on how one defines "extreme poverty." Some identify it as living on less than $1.90 per day, which places some 734 million people—more than the total number of people on Earth prior to the year 1700—in the category. But this cut-off is arbitrary. As Jason Hickel observes,

> the UN's FAO says that 815 million people do not have enough calories to sustain even "minimal" human activity. 1.5 billion are food insecure, and do not have enough calories to sustain "normal" human activity. And 2.1 bil-

> lion suffer from malnutrition. How can there be fewer poor people than hungry and malnourished people? … Lifting people above this line doesn't mean lifting them out of poverty, "extreme" or otherwise.

Third, whatever "progress" humanity may have made with respect to its own desire to maximize economic growth, consume more resources, make money, and so on, our impact on the environment has been nothing short of *catastrophic*. The data here are truly staggering, and much too numerous for the present chapter. (See chapter 7 of my book *The End* for some mind-blowing statistics about how bad the environmental crisis is today.) Suffice it to say that when Ord writes "the track record of technological progress and the environment is at best mixed," he commits the rhetorical crime of prevaricating. It is not in any way "mixed." It is unambiguously horrendous.

Fourth, it's astounding to see someone wax poetic about "progress" in the very same books and papers that identify science and technology as the main reason we face *unprecedented threats to our survival*. In one breath it's "The world is so much better today than ever before!" while in another it's "We now stand closer to the Precipice of total annihilation than we ever have!" The Doomsday Clock, for example, is currently set to 100 seconds before midnight, or doom. This clock was created in 1947, two years into the Atomic Age. But prior to, say, the twentieth century, if the Doomsday Clock had existed, it would have been set to something like 100 seconds *after* midnight on the *same day*. In other words, the minute hand would have been rewound almost 24 hours—that's how *low* the overall risk of extinction was before the twentieth century.

How can anyone think that we've made progress *overall* when the chance of extinction is orders of magnitude higher than *ever before* in our species' history?[24] How are the risks of technology "outweighed by the benefits it has brought" when we stand at the crum-

bling edge of the proverbial Precipice? If human survival were what matters, sane people would be screaming in unison that we need *less* rather than *more* technology.[25] But survival matters to longtermists only as a means to the end of maximizing impersonal value, value, value. This is why proceeding through the ever-more labyrinthine obstacle course of existential hazards before us is worth the risk, for them, of total human annihilation. The more influential longtermism becomes, the harder it will be for the rest of us to change this.

Chapter 9: Conclusions

"So you want to save the world. As it turns out, the world cannot be saved by caped crusaders with great strength and the power of flight. No, the world must be saved by mathematicians, computer scientists, and philosophers."

These are the words of Luke Muehlhauser, currently a research analyst for the EA-aligned Open Philanthropy Project, which has given tens of thousands of dollars to organizations engaged in longtermist projects and research.[26] Silly as they are, they capture, I believe, the grandiosity of longtermist thinking. Longtermists are the prophets and messiahs of the Precipice, out to "save the world" through more science and technology, ultimately leading to the Promised Land of technological maturity. The goal? Saturating our future light cone with intrinsic value by colonizing space, subjugating nature, maximizing economic productivity, simulating huge numbers of conscious beings, and so on.

There are reasons to worry that this worldview is an information hazard. If it were to become influential among politicians or the public, it could precipitate all sorts of harms done in the name of the "greater cosmic good." This is a dangerous, millennialist ideology according to which the means justify the ends and the end is, in Bostrom's canonical formulation, nothing more or less than *Utopia itself*.

More than anything, I want this mini-book to help rehabilitate "longtermism," and hence Existential Risk Studies. As stated above, we very much *do* need more sober reflection of, and strategic thinking about, the future of humanity. We live in a fragile, myopic society confronting slow-motion catastrophes like climate change and the sixth mass extinction that threaten the continued existence of this society. Please do care about the long-term—but don't be a longtermist.[27]

Version 1.1

- This mini-book initially stated “In contrast, because of technology, the probability of total human annihilation according to *longtermists themselves* hovers around 30 percent *this century*.” The figure should be “16.6 [Ord’s estimate] and 20 percent.”

[1] I put "intellectuals" in scare quotes because Sam Harris has propagated a considerable number of unscientific claims about race, IQ, feminism, and other topics. For an amusing collection of some fairly horrendous statements by Harris, click here.
[2] The word "longtermism" has been around for a while, but the sense employed by Effective Altruists (EAs) is novel. Consequently, there is ongoing debate about how exactly it should be defined. EA philosophers have distinguished between, for example, "longtermism," "strong longtermism," "very strong longtermism," "axiological strong longtermism," and "deontic strong longtermism." Because the meaning of the term remains unsettled, "longtermism" is a moving target. It might even be that there are some variants of longtermism that dodge the criticisms leveled below. See also this introductory article by Fin Moorehouse.
[3] For a good critique of one aspect of transhumanism, see "The Irrationality of Transhumanists" by Susan Levin here.
[4] See, for example, section 1 of this *Stanford Encyclopedia of Philosophy* article. Note that utilitarianism is the paradigm case of consequentialism.
[5] Note that virtually no philosophers today are average utilitarians.
[6] For a short introduction to this idea, click here: https://plato.stanford.edu/entries/original-position/#HisBacMorPoiVie.
[7] As Will MacAskill put this very point in a 2018 podcast interview with 80000 Hours: "The third argument [for utilitarianism] is rejecting the idea of personhood, or at least rejecting the idea that who is a person, and the distinction between persons is morally irrelevant. The key thing that utilitarianism does is say that the trade-offs you make within a life are the same as the trade-offs that you ought to make across lives. I will go to the dentist in order to have a nicer set of teeth, inflicting a harm upon myself, because I don't enjoy the dentist, let's say, in order to have a milder benefit over the rest of my life. You wouldn't say you should inflict the harm of going to the dentist on one person intuitively in order to provide the benefit of having a nicer set of teeth to some other person. That seems weird intuitively. … [O]nce you reject this idea that there's any fundamental moral difference between persons, then the fact that it's permissible for me to make a trade off where I inflict harm on myself now, or benefit myself now in order to perhaps harm Will age 70 … Let's suppose that that's actually good for me overall. Well, I should make just the same trade offs within my own life as I make across lives. It would be okay to harm one person to benefit others. If you grant that, then, you *end up with something that's starting to look pretty similar to utilitarianism*" (italics added).
[8] In contrast, Kant argued that rational beings like us *are* ends in ourselves.
[9] For discussion, see the "Separateness of Persons and Distributive Justice" section of Derek Parfit's book *Reasons and Persons*. Other scholars have attempted to avoid this aspect of total utilitarianism here.
[10] Note that, as of 2013, *only 23 percent* of professional philosophers surveyed preferred consequentialism to deontology, virtue ethics, or other ethical theories. For many philosophers, arguments like those articulated by Bostrom are reasons to *reject* utilitarian ethics because of their patent absurdity.
[11] Note that the *Journal of Transhumanism* is now called the *Journal of Evolution and Technology*.
[12] Bostrom gives mixed signals about whether technological maturity *requires* space colonization to have already happened. For example, he writes that "a technologically mature civilisation could (presumably) engage in large-scale space colonisation through the use of automated self-replicating 'von Neumann probes.'" Yet it is unclear how we could attain "capabilities affording a level of economic productivity and control over nature that is close to *the maximum that could feasibly be achieved*" (italics added) *without* spreading through as much of our future light cone as possible.
[13] In other words, the line of reasoning goes in the opposite directions: from premises to conclusion versus from conclusion to premises.
[14] This is to say, as long as influential people give the view legitimacy by endorsing it, or identifying it as a serious view.

[15] Although Bostrom defines this in terms of "true" information, misguided ideologies like longtermism can constitute hazards as well.
[16] See, for example, (i) through (v) in this paper of mine, and section 2.2 (mislabeled "2.1" in the paper) of my Great Challenges Framework publication.
[17] See the Unilateralist's Curse.
[18] Note that Beckstead attempts to argue that the future is overwhelmingly important without committing to what he later described as "a highly specific view about population ethics (such as total utilitarianism)."
[19] A peculiar community of people who have actually encouraged young people to work on Wall Street so that they can donate large sums of money to charity. As Will MacAskill, one of the most prominent EAs, puts it, "To save the world, don't get a job at a charity; go work on Wall Street." For the record, I think this is very bad advice. See also Amia Srinivasan's excellent critique of the idea here.
[20] For one, so long as technology continues to be developed, there is no reason whatsoever to expect the level of existential risk to stabilize or decline. To the contrary, it is likely to increase.
[21] Bostrom lists a few other probability estimates from other scholars but, oddly, gets them almost entirely wrong. For example, he says that John Leslie "estimated a probability that we will fail to survive the current century: 50 percent. Similarly, the Astronomer Royal [i.e., Lord Martin Rees], whom we heard speak yesterday, also has a 50 percent probability estimate." Leslie's estimate was actually a 30 percent chance of extinction within the next five centuries, and Rees' 50 percent estimate concerned civilizational collapse, not extinction.
[22] Many longtermists believe that once we spread beyond Earth, the total existential risk will sharply decline. The reason is that, just as the probability of extinction is inversely related to the geographical spread of a species (i.e., the more spread out, the less chance that, say, a natural disaster will eliminate the species), the greater our cosmographic spread, the lower the chance that a single catastrophe will terminate our evolutionary lineage. But there are very strong reasons for believing that space colonization could *greatly exacerbate* the risk. The most authoritative account of this view is given in the chapters of Daniel Deudney's book *Dark Skies*. A summary of at least some of the key points can be found in this short article of mine. To date, no space expansionist (i.e., advocate of space colonization) has provided a convincing refutation of these points, so we should assume for the time being that there really is no "Planet B."
[23] Intriguingly, there is one instance in Bostrom's oeuvre in which he explicitly acknowledges that "progress" is the wrong word to use. In this paper, he writes: "It may be tempting to refer to the expansion of technological capacities as 'progress.' But this term 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." Yet he uses the term "technological progress" in subsequent articles.
[24] Worse, Bostrom seems to endorse a nightmarish form of ubiquitous, highly invasive state surveillance of individuals as part of this "preventive policing" proposal. He appears to believe that a "high-tech panopticon" of some sort will be necessary to prevent omnicide given the growing power and accessibility of dual-use emerging technologies.
[25] Or at least right now, for the foreseeable future. We are simply too irresponsible—even for nukes. (It's pure dumb luck that the Cold War never turned hot. There were so many near-misses, for example, that I personally have no doubt that if history were rewound to 1945 and played again just *once or twice*, civilization would not have survived.)
[26] In fact, it gave "two grants totaling $38,350 to the Centre for Effective Altruism (CEA) to support the promotion of Toby Ord's book, *The Precipice: Existential Risk and the Future of Humanity*."

[27] A similar critique by Ben Chugg, in which he argues that "longtermism is a dangerous moral ideal," can be found here. Another critique by Vaden Masrani is here. Note also that this mini-book draws in parts from my forthcoming intellectual history book titled *Human Extinction: A History of Thinking About the End of the World*.