Existential Risks: A PhilosophicalAnalysis

Phil Torres
In *Inquiry: An Interdisciplinary Journal of Philosophy*

Abstract: This paper examines and analyzes five definitions of "existential risk." It tentatively adopts a pluralistic approach according to which the definition that scholars employ should depend upon the particular context of use. More specifically, the notion that existential risks are "risks of human extinction or civilizational collapse" is best when communicating with the public, whereas equating existential risks with a "significant loss of expected value" may be the most effective definition for establishing existential risk studies as a legitimate field of scientific and philosophical inquiry. In making these arguments, the present paper hopes to provide a modicum of clarity to foundational issues relating to the central concept of arguably the most important discussion of our times.

Section 1: Introduction

The field of "existential risk studies" was essentially founded by John Leslie's 1996 book *The End of the World: The Science and Ethics of Human Extinction*. However, it lingered in a "pre-paradigmatic" state for several years until the publication of Nick Bostrom's 2002 paper "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards," which drew heavily from Leslie's tome. Since the mid-2010s, numerous organizations focusing on studying existential risks have sprouted up at universities like Oxford and Cambridge, and multiple books on the topic have been added to the scholarly literature, including *Here Be Dragons: Science, Technology, and the Future of Humanity* (2016) and *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* (2017).

Yet there remains some disagreement, or at least a lack of agreement, about the semantics of the field's central concept: existential risk. The aim of the present paper will be to examine the five most prominent definitions employed by existential risk scholars (section 2) and then to offer some critical remarks about which definitions ought to be used in different conversational settings (section 3). Ultimately, I will endorse a *context-dependent semantic pluralism* whereby the best definition to use depends on the intended audience.[1] The fact is that existential risk studies has an integral activist component: it aims not merely to generate new knowledge about the mind-independent world, but to nontrivially influence the developmental trajectory of human civilization. Thus, it is essential that the ideas and insights outlined in the scholarly literature on existential risks are effectively conveyed to non-experts, including politicians, policymakers, business leaders, and the general voting public. This means that, while an agreed-upon technical vocabulary may economize discussion between experts, it is equally important that non-esoteric language is utilized when discussing the ideas and insights with those not "in the know."

The ultimate ambition of this paper is to offer a clear account of how the term "existential risk" has been used by scholars and some considered thoughts about how it should be employed to best advance the normative aims of the field.

Section 2: Five Definitions of "Existential Risk"

A risk is typically defined as the probability of an undesirable event multiplied by its consequences. The notion of an *existential* risk has nothing to do with the first multiplicand; all that's relevant are the consequences. This section will examine five distinct definitions of the term. These are existen-

tial risks as (i) human extinction, (ii) human extinction or civilizational collapse, (iii) human extinction or a permanent and drastic loss of potential, (iv) any catastrophe with pangenerational-crushing effects, and a significant loss of expected value. Taking these in order:

(i) *Human extinction*: An event X is an existential risk if and only if X could cause the extinction of humanity. This is probably the most common definition in the scholarly literature and popular media. For example, Jason Matheny (2007) defines existential risks as "a subset of catastrophic events—those that could extinguish humanity," and Karim Jebari (2015) asserts that existential risks "threaten the continued existence of mankind." Similarly, the journalist Ross Anderson introduces an interview with Bostrom for *The Atlantic* like this: "Bostrom, who directs Oxford's Future of Humanity Institute, has argued over the course of several papers that *human extinction risks* are poorly understood and, worse still, severely underestimated by society. Some of these *existential risks* are fairly well known, especially the natural ones" (Ross 2012, italics added).

There are several reasons to accept this definition. (a) The standard dictionary defines "existential" as "of or relating to existence; involving or relating to the existence of a thing" (OED 2018), which implies that "an existential risk to humanity" is "a risk to the existence of the species denoted by the word 'humanity.'" So, it aligns with common usage of the term "existential." (b) Relatedly, it corresponds to "a natural division and is easy to understand," as Owen Cotton-Barratt and Toby Ord (2015) observe. Whereas phenomena like *flourishing* are spectral, coming in gradations, extinction is binary: A species is either existing or non-existing, extant or extinct, dead or alive. This makes it a non-vague concept, unlike *flourishing*, and one that is objective in nature, in the sense that the concept is non-axiological. (c) And finally, one could argue that the term "extinction" picks out a natural kind—specifically, what Brian Ellis (2001, 2002) calls a "dynamic kind." It "cuts the world at its joints," as it were, and appears to be the sort of phenomenon that could participate in laws of nature. As it happens, almost no work within the philosophy of biology has examined the notion of species extinctions, although there are no doubt many interesting issues to be dissected here.

But there are also several reasons to reject this definition. Consider the following two scenarios. Scenario 1: Imagine that *Homo sapiens* evolves through anagenesis into a new species of "posthumans." This could occur either naturally or through cyborgization, whereby technological modifications significantly alter ways our core capacities of cognition, healthspan, and emotionality (see Bostrom 2008). The result would be the "phyletic extinction" (or "pseudo-extinction") of *Homo sapiens*, and consequently this transition would instantiate an existential risk on the definition above. But now imagine that these posthumans have expanded intellectual capacities, are free of all diseases, don't senesce, and are able to experience degrees of happiness and pleasure that far surpass what any human is acquainted with. While there are bioconservatives, such as Frances Fukuyama (2002) and Leon Kass (2002), who would still oppose the realization of this posthuman future, most people, I conjecture, would see it as desirable. The problem is clear: If one maintains that, as Bostrom puts it, "the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole," then we should do everything necessary to prevent *Homo sapiens* from becoming *Posthomo supersapiens*. But this conflicts with the intuition that just the opposite should be the case, on certain assumptions.

One solution is to redefine "human" as including what David Roden (2014) refers to as our "wide descendants," which denotes "human, machine, cyborg, etc." Thus, the term "human" would encompass potential posthumans that exhibit the right genealogical and axiological relation to current humans.[2] However, expanding the semantics of "human" to count future events in which our posthuman descendants die out would require extra explication, as most non-experts no doubt associate the term with *Homo sapiens* rather than future species that share an unbroken evolutionary lineage with us

and are related in the right sort of axiological way. In fact, a number of population ethicists have argued that the badness of human extinction isn't just the loss of life caused by the extinction event itself, but the loss of all the future beings who could have come into existence had we survived. Derek Parfit and others have been explicit that the relevant class of beings here includes posthumans; as he writes, "life can be wonderful as well as terrible, and we shall increasingly have the power to make life good. Since human history may be only just beginning, we can expect that future humans, or supra-humans, may achieve some great goods that we cannot now even imagine" (Parfit 2017). Thus, if "the destruction of mankind would be by far the greatest of all conceivable crimes" (Parfit 1984), the reason is partly because, if we survive, we could undergo phyletic extinction by becoming a much superior kind of species that achieves immense goods inscrutable to contemporary minds.

Scenario 2: Whereas the scenario above considers a future in which *Homo sapiens* is anagenetically replaced by something better, now imagine a future in which we develop along a "degenerate trajectory" that yields conditions of profound suffering. Call such an outcome a "hyper-existential risk." Anders Sandberg (2014) appears to gesture at this possibility when he writes that, "while there might be posthuman states of great value, there are also potential existential risks threatening futures with no or extremely negative value. Insofar as we can influence what future we might reach, we may have a far greater moral responsibility than is commonly envisioned because the stakes are higher." A related idea is that of a "suffering risk," which David Althaus and Lukas Gloor define as any event "that would bring about suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far" (Althaus and Gloor 2018). Hyper-existential and suffering risks are distinct because, for example, the latter would include future configurations in which the total amount of pleasure exceeds the total amount of suffering, yet the total amount of suffering is astronomical. It would also seem to exclude situations in which, say, humanity fails to colonize space, its population dwindles to 1 billion, and everyone alive experiences constant torture-level misery until the oceans evaporate. Thus, the badness of hyper-existential risks will almost certainly be less controversial than the badness of suffering risks, which are most compelling if one champions a negative total utilitarian theory (or "suffering-focused ethics") (see Gloor and Mannino 2018).[3] The point is that there could be future scenarios in which going extinct—in which a kind of species euthanization—might be unambiguously desirable. Thus, once again, if one maintains that existential risks should be avoided at all costs, then one would oppose efforts to prevent our extinction even when the alternative is a hyper-existential or suffering catastrophe. This seems counter-intuitive.

In sum, the advantage of this definition is that it is *simple*, but the drawback is that it is *simplistic*. I will have more to say about this in section 3.

(ii) *Human extinction or civilizational collapse*: An event X is an existential risk if and only if X could cause either the extinction of humanity or the collapse of civilization. This is not as common in the scholarly literature, although there are references to civilizational collapse in the relevant contexts. For example, while Lord Martin Rees argues that "the worst possible disaster [is] the foreclosure of intelligent life's future through the extinction of all humankind," he focuses a lot on the survival of civilization—at one point arguing that "the odds are no better than fifty-fifty that our present civilisation on Earth will survive to the end of the present century" (Rees 2003). In fact, the Centre for the Study of Existential Risk (CSER), which Rees cofounded, initially used "risks of extinction" and "existential risks" synonymously, but it now describes itself as "dedicated to the study and mitigation of risks that could lead to human extinction or civilisational collapse."[4]

As with the first definition, this one has the virtue of being easy to grasp. Most of us have a robust intuitive understanding of what constitutes civilization, and thus what its collapse would entail. It also encompasses scenarios of planetary-scale destruction that the first definition ignores. For example,

many scholars would describe anthropogenic climate change as posing an existential risk, but few believe that it will directly bring about our extinction. Rather, it is much more likely to cause civilizational collapse. This suggests that existential risks as human extinction *or* civilizational collapse is better than existential risks as human extinction, since this definition draws attention to the fact that some unfathomably bad disasters could nonetheless be *survivable*.

But there are also some problems with this definition. The most significant is that, one could argue, human extinction and civilizational collapse do not have comparable ethical implications. Since human extinction is irreversible, it would result in the loss of potentially vast amounts of future value. In contrast, the collapse of civilization does not necessarily preclude the realization of such value: Those who survive could rebuild civilization over the course of decades or centuries. Consider Bostrom's (2002) point that "from the perspective of humankind as a whole—even the worst of [catastrophes so far in human history] 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." Thus, the global village could implode without this significantly altering the total amount of value that our lineage creates within our future light cone. Even more, civilizational collapse need not entail a major loss of human life—although it is worth noting that even if 99 percent of the current population were to perish, this would still leave 76 million people alive. For such reasons, human extinction appears to be qualitatively worse than civilizational collapse, however unfathomably horrible the latter would be. The point is this: Resources to mitigate existential risks are finite; thus, we should prioritize their use based on which threats are the most *serious*—holding tractability and neglectedness equal. But if the first disjunct in this definition refers to an outcome that is far more serious than the second disjunct, then including civilizational collapse as one form of existential catastrophe could end up taking resources away from preventing human extinction.

Another issue is that it is possible to realize a hyper-existential risk without causing either human extinction or civilizational collapse. Thus, if one wishes for the term "existential risk" to encompass hyper-existential catastrophes as a special case—specifically, as the worst kinds of existential catastrophes—then these definiens are inadequate. Indeed, it could be that continued technological innovation and civilizational development along certain trajectories actually *enables* the realization of a hyper-existential risk, perhaps in the form of an oppressive totalitarian society in which a majority reside in profoundly squalid conditions of inscrutable misery. This suggests that, if existential risks are supposed to be "worst-case outcomes" for humanity, it is inadequate. Yet, as I will argue below, it may nonetheless be useful in non-academic discussions of existential risk phenomena.

(iii) *Human extinction or a permanent and drastic loss of potential*: An event X is an existential risk if and only if X could cause either the extinction of humanity or a permanent and drastic loss of our potential for desirable future development. Since this definition emerges from the work of Bostrom (2002, 2013), I will call it Bostrom's "*lexicographic definition*" to distinguish it from another definition put forth in the same manuscripts, discussed below. Although perhaps not the most commonly used definition in the literature, it is arguably the most canonical, since it is the first formal definition of the term, propounded in Bostrom's 2002 paper that, once again, more or less officially founded the field.[5] Notice that, like the previous definition, this one is also disjunctive, with each disjunct being sufficient for an existential catastrophe to have occurred.

The crucial question is how one should interpret the value-laden second disjunct. What does "desirable future development" mean? For a Benatarian anti-natalist, the most desirable future would be no future at all—i.e., the voluntary and perhaps "phased" extinction of *Homo sapiens*, although this is in tension with the first disjunct. The same goes for a "strong" negative utilitarian: The most desirable

future development would be the total elimination of the very possibility of sentient life in the universe. In contrast, for an anarcho-primitivist, it would be the creation of a "future primitive" state in which humans live in small-scale egalitarian societies that are sustainably embedded within the ecological system. And so on.

Bostrom himself appears to interpret the second disjunct by combining total utilitarianism and transhumanism. Thus, he has previously argued for prioritizing the colonization of space[6] for the purpose of filling our future light cone with as much value as possible. On the total utilitarian view, one might say that *morality abhors a vacuum*. Furthermore, just as we should maximize the total number of happy people in the universe, we should also use person-engineering technologies to optimize our individual capacities to be happy. This would entail the creation of a new lineage, or of multiple lineages, of posthuman beings who, by definition, have cognitive abilities, healthspans, or emotional ranges that far exceed the best attainable by any organism with a human-type genome (Bostrom 2008). Such considerations lead Bostrom, in his 2002 paper, to interpret the axiological term "potential" as referring to "the transition … from a human to a 'posthuman' society." He is more precise about this in his 2013 follow-up on the topic, in which he identifies the following four existential risk "failure modes" (quoting):

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

In every case, existential risks are characterized as failures to establish a stable state of technological maturity. By "technological maturity," Bostrom means "the attainment of capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved" (Bostrom 2013). Why does this matter? Because technological maturity would be the situation most conducive to realizing astronomical amounts of value, in accordance with total utilitarianism and transhumanism. As Bostrom puts it,

> the capabilities of a technologically mature civilisation could be used to produce outcomes that would plausibly be of great value, such as astronomical numbers of extremely long and fulfilling lives. More specifically, mature technology would enable a far more efficient use of basic natural resources (such as matter, energy, space, time, and negentropy) for the creation of value than is possible with less advanced technology. And mature technology would allow the harvesting (through space colonisation) of far more of these resources than is possible with technology whose reach is limited to Earth and its immediate neighbourhood (Bostrom 2013).

It follows from these observations that one could remove the correlative conjunction in the lexicographic definition and reconstruct it more succinctly as follows: "*An event X is an existential risk if and only if X could permanently prevent humanity, broadly defined to include some possible posthuman descendants, from attaining a stable state of technological maturity*." This is the essence of Bostrom's first definition: the ultimate goal of

civilizational development should be producing mature technology in a sustainable manner, so any event that prevents this from happening must be avoided at all costs.

This definition handles Scenario 1 from above without difficulty. As Bostrom (2013) writes, "the permanent foreclosure of any possibility of this kind of transformative change of human biological nature may itself constitute an existential catastrophe." Bostrom also suggests that all four scenarios—and thus both disjuncts—are "of comparable seriousness, entailing potentially similarly enormous losses of expected value." Once again, this contrasts with the second definition in which the second disjunct has far less severe moral implications than the first disjunct, given certain assumptions.

But there are several problems with this definition as well. First, consequentialism is a minority view among philosophers, with only 23.6 percent accepting it as of 2013 (Bourget and Chalmers 2014). And, of course, not all consequentialists are utilitarians, just as not all utilitarians accept the aggregative and impersonal perspectives of the total view. Rather, some hold that there is a diminishing marginal value to creating new people, that the value of future lives is less than the value of current lives (temporal discounting), or that failing to create extra people isn't bad even if those people would live extremely happy lives, since acts can only be bad if they are *bad for* someone (the person-affecting intuition). According to the latter view, Parfit is wrong to argue that the difference between 99 percent and 100 percent of humanity dying out is far greater than the difference between 1 percent and 99 percent dying out. Since the failure to bring into existence a potentially astronomical number of future people with worthwhile lives harms no one, human extinction is worse than the loss of 99 percent of humanity by only one percentage point. Note also here that the total view engenders the Repugnant Conclusion and violates the intuition of "procreational asymmetry," which implies, in slogan form, that we should focus on making people happy rather than making happy people.

Second, many would likely object to the transhumanist values that Bostrom imports into his definition. Bioconservatives, for example, oppose "enhancive" modifications to the human phenotype, especially modifications that would alter our "human essence"—what Francis Fukuyama (2002) terms "Factor X"—upon which liberal democracies are founded. But even some techno-progressives might find the emphasis on technological maturity off-putting: It could be seen as too crass, technocratic, ecologically exploitative, capitalistic, colonialist, and so on. The present author, in fact, knows of several scholars who balk at the lexicographic definition for precisely this reason. This is worrisome because it could lead such individuals to reason as follows: "Since existential risks are scenarios that prevent the attainment of technological maturity, and since I don't care about subjugating nature, maximizing economic productivity, etc., it follows that I must not care about existential risks." The same could be said about the transhumanist desideratum of exploring the posthuman realm: why is failing to radically enhance the human species so undesirable? There are also issues relating to personal identity that arise in this context: it could be that the transformations required to become posthuman result in a genuinely new person, in the metaphysical sense (see author). From this perspective, the true risk is that everyone would essentially commit suicide by augmenting their brains and bodies.

Yet another downside of the lexicographic definition is that it doesn't adequately address Scenario 2 from above. As Althaus and Gloor (2018) observe, this definition

> is unfortunate in that it lumps together events that lead to vast amounts of suffering and events that lead to the extinction … of humanity. However, many value systems would agree that extinction is not the worst possible outcome, and that avoiding large quantities of suffering is of utmost moral importance.

In other words, Bostrom doesn't distinguish between scenarios that could (a) cause our extinction, and (b) result in astronomical amounts of suffering. Thus, one could argue as per above, that this failure is actually quite dangerous given the following conditions: humanity embraces the maxipok rule, has limited resources to mitigate existential risks, and could encounter a hyper-existential or suffering catastrophe. Surely we should—as the maximin (and minipnok[7]) rule implies—at least prioritize avoiding future scenarios that are *worse than extinction* before aiming for a techno-utopian future. Along these line, consider that the following two scenarios fall within the very same moral category on this definition: (1) An asteroid strikes Earth, resulting in an impact winter that kills every human within two dreadful weeks, and (2) our civilization persists at roughly its current level of technological development, with the same number ofpeople having approximately the same quality of life for the next 1 billion years. The latter would be an instance of "permanent stagnation" in Bostrom's quadripartite typology, yet it seems unambiguously better than the former. Surely most people would concur that spending the same amount of resources to avoid (2) as to avoid (1) would be misguided.

A final objection concerns the "temporal permanence" criterion of the second disjunct. As Cotton-Barratt and Ord (2015) write, consider "a totalitarian regime [that] takes control of the earth; there is only a slight chance that humanity will ever escape." Should this count as an existential catastrophe on Bostrom's definition? Perhaps it depends on *when* one answers: in the midst of the regime's global control over civilization, one might suspect at T1 that an existential catastrophe has occurred; but if this regime were to dissolve at a later time T2, then, although there *had been* copious evidence of an existential catastrophe at T1, an existential catastrophe had not *in fact* occurred. In other words, deciding that an event E is an existential catastrophe is an epistemological issue, whereas E actually being an existential catastrophe is a metaphysical one. But this does not completely solve the problem, as Cotton-Barratt and Ord note:

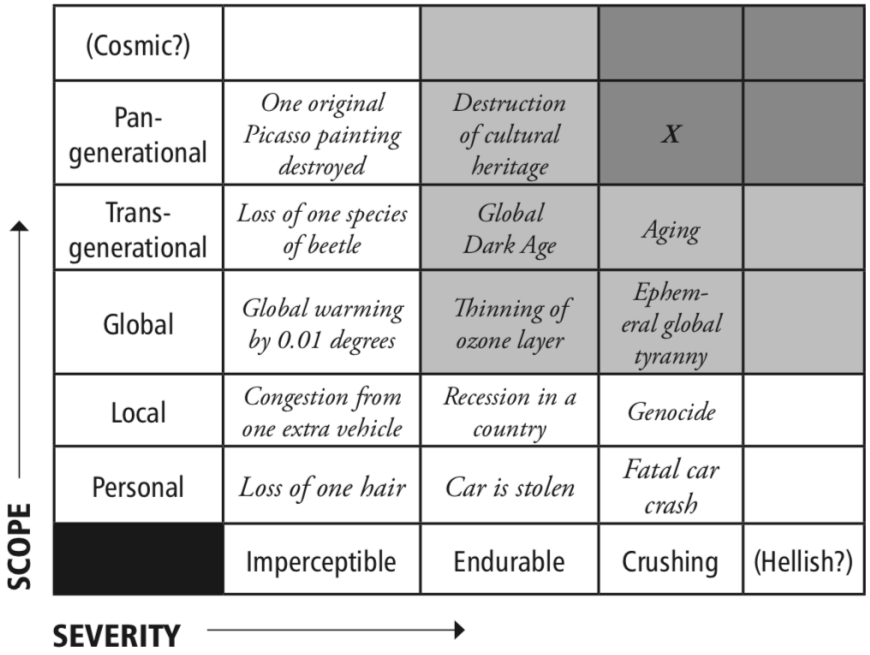
> Bostrom's definition doesn't [convincingly] specify whether it should be considered as one. Either answer leads to some strange conclusions. Saying it's not an existential catastrophe seems wrong as it's exactly the kind of thing that we should strive to avoid for the same reasons we wish to avoid existential catastrophes. Saying it is an existential catastrophe is very odd if humanity does escape and recover—then the loss of potential wasn't permanent after all. The problem here is that potential isn't binary. Entering the regime certainly seems to curtail the potential, but not to eliminate it (Cotton-Barratt and Ord 2015).

Put differently, "potential" isn't binary but "temporal permanence" is, and this poses a problem for the lexicographic definition.

In brief, this definition is more sophisticated than the first two, but its connection to total utilitarianism and transhumanism could make it less attractive to individuals who care about the future of humanity but reject one or the other of these axiological systems. It also appears susceptible to certain counter-examples.

(iv) *A pangenerational-crushing catastrophe*: An event X is an existential risk if and only if X could have undesirable consequences for humanity that are "pangenerational" and "crushing." Let's call this Bostrom's "*typological definition*." First, whereas the lexicographic definition is disjunctive, this one is conjunctive: it specifies two conditions—i.e., being pangenerational and being crushing—that risks must satisfy to count as existential. Second, as the appellation implies, this definition derives from a general typology of risks that Bostrom delineates, and then modifies, in three papers. The typology is based on a particular conceptual analysis of *risk*, according to which a risk is the probability of an undesirable

event multiplied by its consequences, where the consequences of a risk are dissected into two components: *Scope*, or the size of the group affected, and *intensity*, or how bad the effects are for the group.



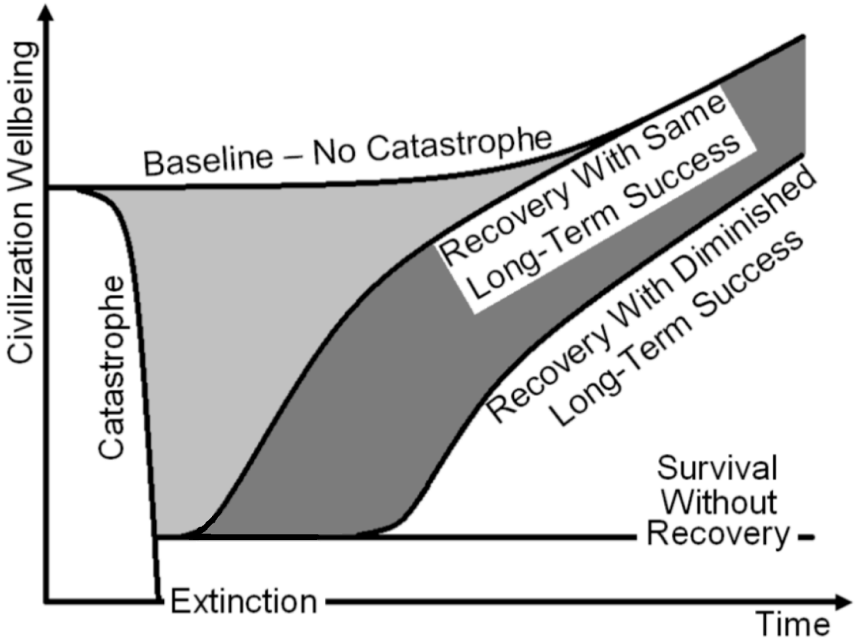
Bostrom's (2013) most recent typology places existential risks in the top right box where the categories of "pangenerational" and "crushing" intersect (see figure 1). The former refers to events "affecting humanity over all, or almost all, future generations," and the latter to events "causing death or a permanent and drastic reduction of quality of life." Thus, Bostrom backs away from the "temporal permanence" criterion of his lexicographic definition, although it is unclear that he recognizes this: while these two definitions are intensionally distinct, they are presumably intended to be extensionally equivalent. If the typological definition includes some events whose effects are permanent and the lexicographic definition does not, then they do not pick out the same class of risky happenings.

There are other problems, though. Consider the categories of the x-axis. First, it seems that "imperceptible" should be a *subcategory* of "endurable," and as such this seems to commit what Gilbert Ryle (1949) famously called a "category mistake." On this logic, one might as well also include as distinct categories "barely noticeable," "moderately annoying," "quite bad," and so on. Furthermore, if some crushing events are by stipulation endurable, then distinguishing "crushing" events from "endurable" events is confused. Similar Rylean issues problematize the y-axis. For example, notice that "personal," local," and "global" are all *spatial* categories whereas "transgenerational" and "pangenerational" are *spatiotemporal* categories. The result is another ontological error.

These problems yield some empirical inadequacies. For instance, consider that a germline mutation, such as the hemoglobin gene mutations that cause sickle-cell anemia, could be restricted to a local or personal scope across space yet will also have transgenerational temporal effects. Yet where in figure 1 might this risk fit? Similarly, Bostrom classifies the "loss of one species of beetle" as a *transgenerational-imperceptible* event. But this overlooks potentially important facts about the geographical spread of the species. If the beetle is restricted to a small region, then its disappearance may indeed be imperceptible, but if its ecological range is more global, then the loss of a single species of beetle could have cantharophilic effects that nontrivially harm the health of larger ecosystems. It may thus be important to

specify whether a transgenerational event of this sort has local or global consequences, yet there is no possibility in this typology of, for example, *local-pangenerational-imperceptible* risks. Finally, at the non-existential risk of caviling, it seems rather odd, rather awkward, to describe aging as a "crushing" risk.

Fortunately, these problems aren't insoluble: (i) One could replace the three/four categories on the x-axis with the two basic categories of "recoverable" and "unrecoverable," thereby more perfectly aligning the typological definition with the lexicographic definition. An "unrecoverable" event would thus be one that forever prevents humanity from attaining a stable state of technological maturity, whereas a "recoverable" event would be one that doesn't. All four existential risk failure modes—i.e., human extinction, recurrent collapse, and so on—would then fall within the unrecoverable category. This would, of course, require that the definition of "pangenerational" changes from "affecting humanity over all, or almost all, future generations" to "affecting humanity for all future generations," but this is how Bostrom seems to think about existential risks in most of his explorations of the topic. Incidentally, the recoverable/unrecoverable distinction nicely maps on to a diagram put forth by Baum et al. 2015, seen in figure 2. As this diagram makes clear, albeit tacitly, an existential catastrophe could have two distinct outcomes, namely, *extinction* and *survival without recovery*.

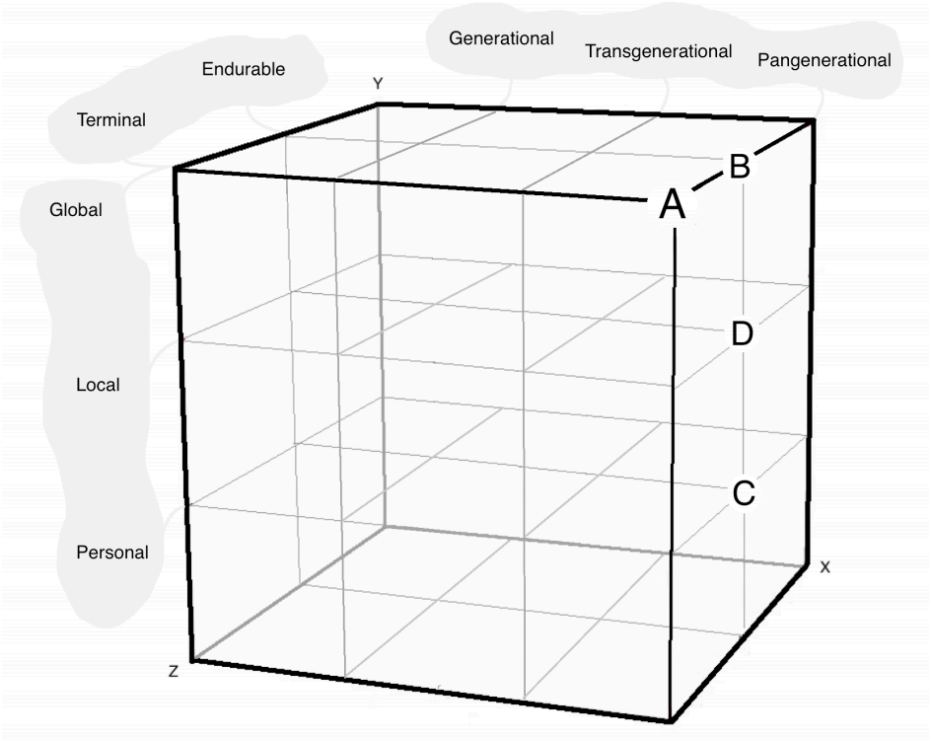


And (ii), one could split the spatiotemporal components of the categories on the y-axis, thus separating this dimension into space and time. The resulting three-dimensional typology would be based on the following augmented conceptual analysis of *risk*:

> A risk is the probability of an undesirable event multiplied by its consequences; the consequences are then analyzed into their scope and intensity, where scope is further decomposed into its spatial and temporal properties. Thus: "risk = [spatial scope, temporal scope, and intensity of consequences] x [probability of occurrence]."

If we revert to an older dichotomy used by Bostrom (2002)—one that distinguishes between "endurable" and "terminal" rather than, as per above, "recoverable" and "unrecoverable" events—this

new typology would classify all existential risks as *global-pangenerational* events, where the first term indicating the risk's spatial scope and the second its temporal scope. Within this conjunctive category, extinction events would constitute *global-pangenerational-terminal* events (node A) and existential risks that



correspond to the second disjunct of Bostrom's lexicographic definition would constitute the subset of *global-pangenerational-endurable* events that irreversibly preclude the stable attainment of technological maturity (node B).[8] In addition to these typological accounts of existential risks, figure 3 also easily accommodates phenomena like aging, hereditary disease (node C), and the loss of a single species of beetle (node D).

While this modified typological definition is more complex, it is also more precise. It is also constant with Thomas Kuhn's (1962) observation that as scientific fields mature, they tend to undergo "a refinement of concepts that increasingly lessens their resemblance to their usual commonsense prototypes." The three-dimensional definition also makes explicit that some existential risks can indeed be survived.

(v) *A significant loss of expected value*: An event X is an existential risk if and only if X could cause the loss of a large fraction of expected value.[9] This definition comes from a short 2015 paper by Cotton-Barratt and Ord, published as a "technical report" by the Future of Humanity Institute (FHI). It is, as they write, primarily motivated by the aforementioned fact that potential isn't binary but permanence is. Recall the totalitarian scenario mentioned when discussing Bostrom's lexicographic definition above; Cotton-Barratt and Ord write that "if we enter into the totalitarian regime and then at a later date the hope of escape is snuffed out, that represents two existential catastrophes under [the expected value] definition." That is to say, humanity "lost most of the expected value when we entered the regime, and then lost most of the remaining expected value when the chance for escape disappeared."

It may be worth pausing on this point for a moment. On the present definition, an existential catastrophe can occur *more than once* in a species' lifetime. The same goes for an existential catastrophe that entails "mere" civilizational collapse, as specified by the second definition above. But this is not the case if one accepts the "existential risks as human extinction," "existential risks as human extinction or a permanent and drastic loss of potential," or "existential risks as pangenerational-crushing events." In all of these cases, humanity has one, and only one, shot at avoiding disaster. Indeed, this is why Bostrom (2002) emphasizes that we must replace our normal reactive approach to dealing with risks with an entirely proactive approach: if even a *single* existential catastrophe were to occur, the game would be over and humanity would have lost (its chance to "fulfill its potential"). Thus, Cotton-Barratt and Ord's definition dilutes this implication of preemptive urgency at least a little by allowing for the possibility that humanity *recovers* from one or more past existential catastrophes. This is a notable difference with respect to the other definitions so far explicated.

Other features of this definition are: first, one can interpret it as a refinement of the second disjunct of Bostrom's lexicographic definition that also discards the first disjunct (as conceptually unnecessary within the definiens). The result is a reorganization of Bostrom's four failure modes of "human extinction," "permanent stagnation," "flawed realization," and "subsequent ruination" under a single conceptual heading, rather than placing the first failure mode under the first disjunct of the lexicographic definition and the last three under the second disjunct. Put differently, it looks arbitrary and inconsistent that Bostrom's definition singles out human extinction—one of four particular failure modes—but then hides the other three failure modes behind the phrase "permanently and drastically compromising our potential for desirable future development." The critical point is that all of thee failure modes are bad because they entail huge losses of expected value within our future light cone.

Cotton-Barratt and Ord's definition also foregrounds an idea—i.e., *expected value*—that nebulously lurks behind Bostrom's discussions of existential risk, making occasional cameos but then returning to the philosophical background. For example, in a discussion about the moral force of the maxipok rule, Bostrom writes that (partly quoted above) "these considerations suggest 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" (Bostrom 2013).

The point is that Cotton-Barratt and Ord's definition could be seen as a more distilled version of Bostrom's definitions, which leads to the next point: third, the expected value definition does not explicitly mention technological maturity, and thus it does not engender the problem of potential allies ignoring existential risk studies because its central aims are "too transhumanistic." For Cotton-Barratt and Ord (2015), the term "'value' refers simply to whatever it is we care about and want in the world"—which, once more, connects with the following point: fourth, this flexibility opens the door for a wide range of axiological interpretations of the definition. As mentioned above, most existential risk scholars appear to espouse total utilitarianism and many are also sympathetic with the transhumanist project; for such scholars, "value" refers to the well-being of sentient life (which ought to be maximized) and "having the opportunity to explore the transhuman and posthuman realms," as Bostrom (2005) puts it.

Another cluster of issues to note is: (a) The first two definitions above specifically reference scenarios that we should avoid, without saying anything about which scenarios we should pursue. On these accounts, the worst-case outcome for humanity would be to *fall victim* to such bad scenarios; the same could be said of the fourth, typological definition. (b) The third, lexicographic definition focuses on what we should *attain*, defining any event that prevents the attainment of technological maturity—a techno-utopian condition of endless possibilities (see Bostrom 2008)—as the worst-case outcome for humanity.

Or, as Cotton-Barratt and Ord (2015) put it, "under Bostrom's definition we are comparing ourselves to the most optimistic potential we could reach."

In contrast, (c) the present definition focuses on avoiding a significant loss of expected value but does not explicitly identify any particular scenario, such as human extinction, civilizational collapse, or failing to reach technological maturity, that humanity must dodge at all costs. Yet Cotton-Barratt and Ord argue that their conception of existential risks gives rise to the mirror concept of what they call "existential hope." This arises from the possibility of (to borrow a term from J.R.R. Tolkien) "existential eucatastrophes," which are the mirror phenomena of existential catastrophes, that could cause "there to be much more expected value after the event than before." Thus, they argue that just as we should *avoid* existential catastrophes, we should *seek out* existential eucatastrophes, examples being the emergence of the first living organisms, any successful dodging of a Great Filter, or, perhaps, the creation of a value-aligned machine superintelligence. From a pragmatic point of view, the idea of existential hope, as Max Tegmark (2018) points out, usefully encourages us to imagine "positive futures not only for ourselves, but also for society and for humanity itself," an activity that may help to overcome the occasional paralysis or defeatism that some people experience when cogitating the topic of existential risks (see also Pinker 2018). In a phrase, the combination of existential risk and existential hope offers a more complete and comprehensive world-picture that humanity could employ when deciding which development trajectories we ought to pursue, and this makes it appealing as an organizing principle for the field of existential risk studies.[10]

But there are also some ostensive problems with this definition. First, as implied above, it may be *too* flexible. As Cotton-Barratt and Ord note, the term "value" is doing much of the work in the definition and, as we have seen, there is a huge range of distinct and mutually exclusive value theories on the marketplace of ideas. Indeed, there is nothing conceptually incoherent about, say, an anarcho-primitivist exclaiming, "We must do everything we possibly can to mitigate existential risks! Follow the max-ipok rule!," but meaning that we ought to destroy the techno-industrial system upon which modern civilization is founded. Second, if one accepts the "astronomical value thesis," according to which the value of the distant future could be unimaginably great (see author), that future lives are worth the same as present lives, and so on, the result could be what Steven Pinker (2011) terms "a pernicious utilitarian calculus" that encourages people to, as it were, redirect the trolly to sacrifice the lives of current humans for the purpose of inflating the probability that innumerably more posthumans come to exist in the future. This is a criticism that applies no less to Bostrom's (utilitarian) conception of existential risks, as Olle Häggström argues. To quote Häggström 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, 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.

And finally, if one interprets value in Cotton-Barratt and Ord's definition as well-being, and well-being in hedonistic terms, then, as David Pearce (2014) observes, total utilitarians are "obliged to convert your matter and energy into pure utilitronium, erasing you, your memories, and indeed human civilisation."[11] A couple of points: (a) One can view this as a *reductio ad absurdum* of hedonistic monism, along the lines of Robert Nozick's (1974) "experience machine" thought experiment, and (b) many people would no doubt argue that converting the universe into utilitronium would yield if not a worst-case outcome, then certainly not the best-case outcome. Thus, the combination of Cotton-Barratt and Ord's general framework plus a widely accepted and respectable moral theory could have potentially devastating consequences, or so one might plausibly argue.

Section 3: Evaluating Definitions

Consistent with Kuhn's observation in section 1, Steven Pinker (2014) observes that esoteric vocabularies are "unavoidable because of the abstractness and complexity of [academic] subject matter[s]." Consequently,

> every human pastime—music, cooking, sports, art—develops an argot to spare its enthusiasts from having to use a long-winded description every time they refer to a familiar concept in one another's company. It would be tedious for a biologist to spell out the meaning of the term *transcription factor* every time she used it, and so we should not expect the tête-à-tête among professionals to be easily understood by amateurs.[12]

Yet the aim of existential risk studies isn't just to further expand the edifice of human knowledge, but—as alluded to in section 1—to communicate these ideas and insights to politicians, policymakers, business leaders, and to all others with the power of the ballot. Informing the public about the ideas and insights generated by existential risk research is integral to the broader program; there is simply no point to devising effective strategies for navigating the obstacle course of threats before us if those strategies will never be implemented by governments, international organizations, corporations, and other relevant entities. To paraphrase Karl Marx's now-hackneyed exhortation (borrowed from Voltaire), *the point is to change the world*. Indeed, a number of leading scholars have made a number of notable media appearances to discuss existential risks and related issues. Essential to the success of the field are "ambassadors of science"—very often, thus far, active contributors to the technical literature—capable of interfacing between the experts and non-experts.

Notice here that different risks require different outreach strategies. For example, humanity is unlikely to prevent further anthropogenic climate change without the public putting significant pressure on political leaders. In democratic states, this means persuading the electorate to vote for candidates who vow to support legislation that reduces our collective carbon footprint, encourages the development of alternative energy sources, and so on—essentially, to resolve the tragedy of the commons. Thus, efforts to convey information about the potentially catastrophic effects of climate change should target the general public. This contrasts with risks like asteroid impacts and, arguably, value-misaligned machine superintelligence. In the former case, public knowledge about near-Earth objects

(NEOs) is mostly irrelevant to the detection and deflection of asteroids and comets that are on a crash course with Earth. So long as NASA and similar organizations around the world remain sufficiently well-funded, the extent to which humanity is collectively safe from a catastrophic collision is not a function of the degree of public informedness. Similar points could be made about superintelligence: On the one hand, avoiding an AI arms race with China or Russia might depend on adequate pressure on politicians by an electorate worried about the possible consequences of another "Cold War," except far less stable given the nature of AI and the larger number of actors that could be involved (Armstrong et al. 2016; author). Thus, on the other hand, when it comes to superintelligence, a positive outcome will depend almost entirely on experts: moral philosophers and computer scientists working on the "value-alignment problem." Alerting the public of the dangers posed by machine superintelligence is unlikely to directly alter progress on determine what our "human values" are and how to define these "in primitives such as mathematical operators and addresses pointing to the contents of individual memory registers" (Bostrom 2014).

Thus, whereas existential risk scholars who are focused on environmental degradation need to engage with the general public, those working on issues like superintelligence should direct their outreach efforts at theoreticians, industry scientists, CEOs, and so on. If this picture is accurate, then it might be optimal to employ different definitions depending on the conversational context. Indeed, this is the position that I will here defend, which I call "context-dependent semantic pluralism." In my view, the second definition of existential risk as either human extinction or civilizational collapse is preferable to the first, and the best definition for communicating to the general public. The reasons are as follows: first, if the question is "What ought our global priorities be?," then the first definition is much too narrow. Although the moral implications of civilizational collapse are less significant than those of extinction, it is useful to at least *signal* to the public that scholars aren't *merely* worried about human extinction—a planetary-scale catastrophe could catapult the human species back into the Stone Age—even if the signal itself is inexact. Indeed, a number of scenarios in the foreground of researchers' attention, such as climate change, seem unlikely to directly cause human extinction but quite likely to seriously damage human civilization.

Even more, civilizational collapse may be more vivid to the average mind than human extinction, in part because few apocalyptic Hollywood movies focus on extinction rather than collapse. Although scope neglect and "psychophysical numbing" could apply to both scenarios, it may be especially applicable to annihilation, thus dulling the emotional impact of warnings of the form, "If X happens, humanity could go extinct." A recent unpublished study, in fact, found that Derek Parfit was right in his conjecture that most people don't see the difference between 99 percent and 100 percent of humanity dying out as much greater than the difference between peace and 99 percent of humanity dying out. Yet, when researchers prodded subjects to think explicitly about the far future and all the good that it could envelop, they shifted their opinion toward Parfit's own view.

With respect to the lexicographic, typological, and expected value definitions of "existential risk," there is a good case that these are too esoteric for a general audience. Bostrom's definition requires some knowledge of moral philosophy and transhumanism, and few members of the public are conversant with decision theory and the technical concept of *expected value*. In sum, the second definition avoids the simplism of the first definition as well as the esoterica of the last three; it constitutes a middle ground between these two ends. I thus recommend its use when discussing existential risks and related topics with lay audiences.

There remains the second question of which definition ought to be the canonical one used by scholars in conversations amongst themselves and with professional scientists and philosophers outside

of the field. Here the simplism of the first definition and imprecision of the second quickly disqualifies them from serious consideration. That is to say, it would be misleading to equate existential risks with either mere human extinction or civilizational collapse; careful axiological reflection implies that what really matters are disasters with long-lasting effects that drastically compromise our *potential* to realize our various individual and collective ambitions for happiness, knowledge, peace, moral growth, and so on. Both of Bostrom's definitions gesture at this idea, which makes them appealing, although as previously noted even some techno-progressives could find the focus on "technological maturity" to be objectionable. Indeed, transhumanism remains a controversial normative-futurological position, and recall from above that a relatively small minority of professional philosophers espouse total utilitarianism. For reasons such as these, I believe that the field of existential risk studies would generally benefit from eschewing both of Bostrom's definitions.

This leaves the fifth definition from Cotton-Barratt and Ord, which I believe, with modification, is superior to all other definitions in the context of tête-à-tête communication *and* efforts to evangelize for the field to scientists and philosophers unfamiliar with its aims, methodologies, and so on. Most people value things in the world and, as Samuel Scheffler (2007) observes, "what would it mean to value things but, in general, to see no reason of any kind to sustain them or retain them or preserve them or extend them into the future?" The ambiguity of "value" in the fifth definition enables a wide range of interpretations, which could result in individuals who value seemingly disparate things—e.g., sports, literature, philosophy, science, the wilderness, standup comedy shows, and 3D movies—to unify behind the single *ultra-goal* of mitigating existential risk. Indeed, it may be an improvement of Cotton-Barratt and Ord's definition to make explicit that "value" can be understood as denoting whatever one happens to care about in this strange place called "the universe." Thus, when a scientist or philosopher outside the field asks what existential risks are, one could respond: An event X is an existential risk if and only if X could cause the loss of a large fraction of expected value, where you (the interlocutor) can define "value" however you'd like. It is my considered opinion that this constitutes the best definition for the purpose of further establishing existential risk studies as a legitimate field of scientific and philosophical inquiry.

The implication of this admittedly brief discussion of context-dependent semantic pluralism is that existential risk scholars will have to pivot from one use to the other in response to different audiences.[13] Given that existential risk studies is an "outward-facing" (or integrally activist) field that strives to bring about nontrivial positive changes in the real world, terminological flexibility may be an essential skill for existential risk scholars to develop.

Section 4: Conclusion

This paper falls within an emerging subfield that I call "the conceptual foundations of existential risk studies." Given the dearth of research so far conducted on the relevant topics, I do not see this paper as the final word. Rather, my goal is much more modest: To merely stimulate further philosophical analyses of this concept by bringing together, for the first time in a single paper, the various definitions that have so far been outlined and discussed seriously in the scholarly literature. Future research on this topic might thus focus on devising better necessary and sufficient conditions for the definiendum, exploring how different theories of concepts (e.g., the "prototype theory," "theory theory") might solve some of the problems specified above, examining additional novel strategies for making the idea of *existential risk* more appealing those outside of the field's perimeter, and communicating important research

results to the public—a necessary step with respect to certain risks like climate change. Thus, my closing remark is an invitation: Much work remains to be done.

Acknowledgements: Many sincere thanks to Simon Beard for detailed comments and criticisms that greatly improved this paper.

References:

Althaus, David, and Lukas Gloor. 2018. Reducing Risks of Astronomical Suffering: A Neglected Priority. Foundational Research Institute. https://foundational-research.org/reducing-risks-of-astronomical-suffer- ing-a-neglected-priority/.

Anderson, Ross. 2012. We're Underestimating the Risk of Human Extinction. *The Atlantic*. https://www.theatlantic.com/technology/archive/2012/03/were-underestimating-the-risk-of-human-extinction/ 253821/.

Armstrong, Stuart, Nick Bostrom, Carl Shulman. 2016. Racing to the Precipice: A Model of Artificial Intelligence Development. *AI & Society*. 31(2): 201-206.

Baum, Seth. 2010. Is Humanity Doomed? Insights from Astrobiology. *Sustainability*. 2(2): 591-603.

Baum, Seth, David Denkenberger, and Jacob Haqq-Misra. 2015. Isolated Refuges for Surviving Global Catastrophes. *Futures*. 74: 45–56.

Baum, Seth, and Itsuki Handoh. 2014. Integrating the Planetary Boundaries of Global Catastrophic Risk Paradigms. *Ecological Economics*. 107: 13-21.

Baum, Seth, Timothy Maher, and Jacob Haqq-Misra. 2013. Double Catastrophe: Intermittent Stratospheric Geoengineering Induced by Societal Collapse. *Environment Systems and Decisions*. 33(1): 168-180.

Beckstead, Nick. 2013. On the Overwhelming Importance of Shaping the Far Future. Dissertation. https:// rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/.

Bourget, David, and David Chalmers. 2014. What Do Philosophers Believe? Philosophical Studies. 170(3): 465-500.

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

Bostrom, Nick. 2005. A Philosophical Quest for Our Biggest Problems. TED. https://www.ted.com/talks/nick_bostrom_on_our_biggest_problems/transcript.

Bostrom, Nick. 2005. Transhumanist Values. *Review of Contemporary Philosophy*. 4. https://nick-bostrom.-com/ethics/values.html.

Bostrom, Nick. 2009. The Future of Humanity. *Geopolitics, history, and International Relations*. 1(2): 41- 78.

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

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

Bostrom, Nick, and Milan Ćirković. 2008. *Global Catastrophic Risks*. Oxford, UK: Oxford University Press.

Carey, Susan. 1985. *Conceptual Change in Childhood*. Cambridge, MA: MIT Press.

Chalmers, David. 2011. Verbal Disputes. *The Philosophical Review*. 120(4): 515-566.

Ćirković, Milan. 2012. Small Theories and Large Risks—Is Risk Analysis Relevant for Epistemology? *Risk Analysis*. 32(11): 1994-2004.

Coyne, Jerry. 2019. A Response from Steve Pinker. *Why Evolution is True*. https://whyevolutionistrue.-wordpress.com/2019/01/29/a-response-from-steve-pinker-to-salons-hit-piece-on-enlightenment-now/.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Ellis, Brian, 2001. *Scientific Essentialism, Cambridge Studies in Philosophy*. Cambridge, UK: Cambridge University Press.

Ellis, Brian. 2002. *The Philosophy of Nature*. Chesham: Acumen.

Ereshefsky, Marc. 2017. Species. *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.). https://plato.stanford.edu/archives/fall2017/entries/species/

Farquhar, Sebastian, John Halstead, Owen Cotton-Barratt, Stefan Schubert, Haydn Belfield, and Andrew Snyder-Beattie. 2017. Existential Risk: Diplomacy and Governance. Global Priorities Project. https:// www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf.

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

FLI. 2018. Existential Risk. Future of Life Institute. https://futureoflife.org/background/existential-risk/.

Fukuyama, Francis. 2002. *Our Posthuman Future: Consequences of the Biotechnology Revolution*. New York, NY: Farrar Straus & Giroux.

Garrick, John. 2017. First International Colloquium on Catastrophic and Existential Risk. *The B. John Garrick Institute for the Risk Sciences*. https://docplayer.net/87926854-First-international-colloquium-on-catastrophic-and-existential-risk.html.

Gloor, Lukas and Adriano Mannino. 2018. The Case for Suffering-Focused Ethics. Foundational Research Institute. https://foundational-research.org/the-case-for-suffering-focused-ethics/.

Glover, Jonathan. 1977. *Causing Death and Saving Lives*. London: Penguin Books.

Goldin, Ian. 2011. Global Shocks, Global Solutions: Meeting 21st-Century Challenges. In David Held, Marika Theros, and Angus Fane-Harvey (eds.), *The Governance of Climate Change*. Cambridge, Uk: Polity Press.

Haslanger, Sally. 2000. Gender and Race: (What) Are They? (What) Do We Want Them To Be? *Nous*. 43(1): 31-55.

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

Holm, Søren. 2017. The Bioethicist Who Cried "Synthetic Biology": An Analysis of the Function of Bioterrorism Predictions in Bioethics. *Cambridge Quarterly of Healthcare*. 26(2): 230-238.

Huntington, Samuel. 1993. The Clash of Civilizations? *Foreign Affairs*. https://www.foreignaffairs.com/articles/united-states/1993-06-01/clash-civilizations.

Jebari, Karim. 2015. Existential Risks: Exploring a Robust Risk Reduction Strategy. *Science and Engineering Ethics*. 21(3): 541-554.

Kass, Leon. 2002. *Life, Liberty, and the Defense of Dignity: The Challenge for Bioethics*. San Francisco: Encounter Books.

Kuhn, Thomas. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Lakoff, George, and Mark Johnson. 1980. *Metaphors We Live By*. Chicago, IL: University of Chicago Press.

Matheny, Jason. 2007. Reducing the Risk of Human Extinction. *Risk Analysis*. 5: 1335-1344.

Millett, Piers and Andrew Snyder-Beattie. 2017. Existential Risk and Cost-Effective Biosecurity. *Health Security*. 15(4): 373-383.

Nozick, Robert. 1974. *Anarchy, state, and utopia*. New York: Basic Books.

Pamlin, Denis, and Stuart Armstrong. 2015. 12 Risks that Threaten Human Civilisation. Global Challenges Foundation. https://api.globalchallenges.org/static/wp-content/ uploads/12-Risks-with-in nite-impact.pdf.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Oxford University Press.

Parfit, Derek. 2017. *On What Matters: Volume Three*. Oxford: Oxford University Press.

Pinker, Steven. 2014. Why Academics Stink at Writing. *The Chronicle Review*. http://wiki.lib.sun.ac.za/images/9/91/2104-pinker-writing.pdf.

Pinker, Steven. 2018. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. New York, NY: Penguin Books.

Price, Huw. 2013. Cambridge, Cabs, and Copenhagen: My Route to Existential Risk. *The Stone, New York Times*. https://opinionator.blogs.nytimes.com/2013/01/27/cambridge-cabs-and-copenhagen-my- route-to-existential-risk/.

Rees, Martin. 2003 *Our Final Hour: A Scientist's Warning*. New York, NY: Basic Books.

Reiss, Julian, and Jan Sprenger. 2017. Scientific Objectivity. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=scientific-objectivity.

Rockström, Johan. 2015. Bounding the Planetary Future: Why We Need a Great Transition. *Great Transition Initiative*. http://www.greattransition.org/publication/bounding-the-planetary-future-why-we-need-a-great-transition.

Ryle, Gilbert. 1949. *The Concept of Mind*. Chicago, IL: University of Chicago Press.

Sagan, Carl. 1983. Nuclear War and Climatic Catastrophe: Some Policy Implications. Foreign Affairs. 62(2): 57-92.

Sandberg, Anders. 2014. The Five Biggest Threats to Human Existence. The Conversation. https://the-conversation.com/the-five-biggest-threats-to-human-existence-27053.

Sandberg, Anders. 2015. Transhumanism and the Meaning of Life. In Calvin Mercer and Tracy Trothen (eds.), *Religion and Transhumanism: The Unknown Future of Human Enhancement*. Oxford, UK: Praeger.

Scheffler, Samuel. 2016. *Death and the Afterlife*. Oxford: Oxford University Press.

Scheffler, Samuel. 2018. *Why Worry About Future Generations?* Oxford: Oxford University Press.

Tegmark, Max. 2018. *Life 3.0: Being Human in the Age of Artificial Intelligence*. New York, NY: Vintage.

Tegmark, Max, and Nick Bostrom. 2005. How Unlikely Is a Doomsday Catastrophe? arXiv.org. https://arxiv.org/pdf/astro-ph/0512204.pdf.

[1] There is, as of this writing, a burgeoning literature on pluralism within the field of philosophy. Of particular interest is Chalmers (2011), especially a particular version of the "method of elimination" that Chalmers calls the "subscript gambit." (Thanks to an anonymous reviewer for calling my attention to this paper.) I draw from this literature, although mostly implicitly.

[2] The axiological constraint here aims to preclude scenarios in which current humans are supplanted by a posthuman population whose moral and, more generally, value commitments are different than ours in some problematic way. As Bostrom (2013) puts this point, "even if another intelligent species were to evolve to take our place, there is no guarantee that the successor species would sufficiently instantiate qualities that we have reason to value. Intelligence may be necessary for the realisation of our future potential for desirable development, but it is not sufficient."

[3] Incidentally, Bostrom appears to overlook these issues in one section of his 2013 paper about existential risks. In defending his maxipok rule, Bostrom argues against the Rawlsian "maximin" rule, which states that, under ignorance, one should pick the action with the best worst-case outcome. Since it is impossible to completely rule out the possibility of extinction no matter which actions we choose, Bostrom claims that "the use of maximin in the present context would entail choosing the action that has the greatest benefit under the assumption of impending extinction," which would of course be Bacchanalian revelry (Bostrom 2013). This is confusing, though, since the maximin rule itself is indifferent to actions that have the same worst possible outcome; thus, decision theorists have devised the "leximin" rule, which states that if two actions have equally bad worst possible outcomes, one should choose the action whose second worst possible outcome is the best. Thus, if one takes extinction to be the worst possible outcome—as Bostrom here does—then one should choose actions that have the best worst possible outcomes next to extinction. Either way, the point is that Bostrom's assumption is clearly wrong: There are many forms of survival, whether in a human or posthuman state, that are far worst than extinction. It is such considerations that lead Althaus and Gloor (2018) to hortatorily declare: "Rather than focusing exclusively on ensuring that there will be a future, we recommend interventions that improve the future's overall quality."

[4] Furthermore, in a discussion of natural global pandemics, Häggström writes that not only do pandemics pose a "global threat," but it may "also pose a risk on the next level, threatening to end civilization as we know it, or even the existence of humanity" (Häggström 2016).

[5] One finds myriad variations of this definition in the literature; for example, the Future of Life Institute (FLI) states that "an existential risk is any risk that has the potential to eliminate all of humanity or, at the very least, kill large swaths of the global population, leaving the survivors without sufficient means to rebuild society to current standards of living" (FLI 2018). Similarly, a Global Challenges Foundation report from 2015 claims to introduce "a new category of global risk," namely, "risks with potentially infinite impact." The report defines a risk of this sort as "where civilisation collapses to a state of great suffering and does not recover, or a situation where all human life ends." This comes close to definition (ii) above, but the emphasis on civilizational collapse events that have irreversible consequences positions it closer to Bostrom's lexicographic definition. Indeed, the authors of the report write that

> the idea to establish a well defined category of risks that focus on risks with a potentially infinite impact that can be used as a practical tool by policy makers is partly inspired by Nick Bostrom's philosophical work and his introduction of a risk taxonomy that includes an academic category called "existential risks." … The infinite category is a smaller subset of "existential risk" and this new category is meant to be used as a tool.

Finally, one of the present authors has, on numerous occasions, somewhat extravagantly limned the concept of *existential risk* as "any future event that either trips our species into the eternal grave of extinction or irreversibly catapults us back into the Stone Age."

[6] Just so long as doing so doesn't interfere with priority number one, namely, the reduction of existential risk.

[7] See author.

[8] Note: the "permanence" condition of the second disjunct is embedded within the concept of *pangenerational*.

[9] Quoted in Cotton-Barratt and Ord 2015.

[10] Some might argue that "existential despair" better contrasts with "existential hope," although Cotton- Barratt argues that this point isn't "necessarily enough to sink the [original] term" of "existential hope" (Cotton-Barratt 2015).

[11] Note that total utilitarianism in particular, and aggregative consequentialism in general, are also beset by what Bostrom (2011) calls "infinitarian paralysis."

[12] This is, of course, a comical reference to Pinker's comments in Coyne 2019.

[13] There are also parallels between this account and Sally Haslanger's (2000) assertion that, "in defining our terms, we must keep clearly in mind our political aims both in analyzing the past and present, and in envisioning alternative futures." We should, as she implores, "begin by asking both in the theoretical and political sense what, if anything, we want [our concepts] to be."