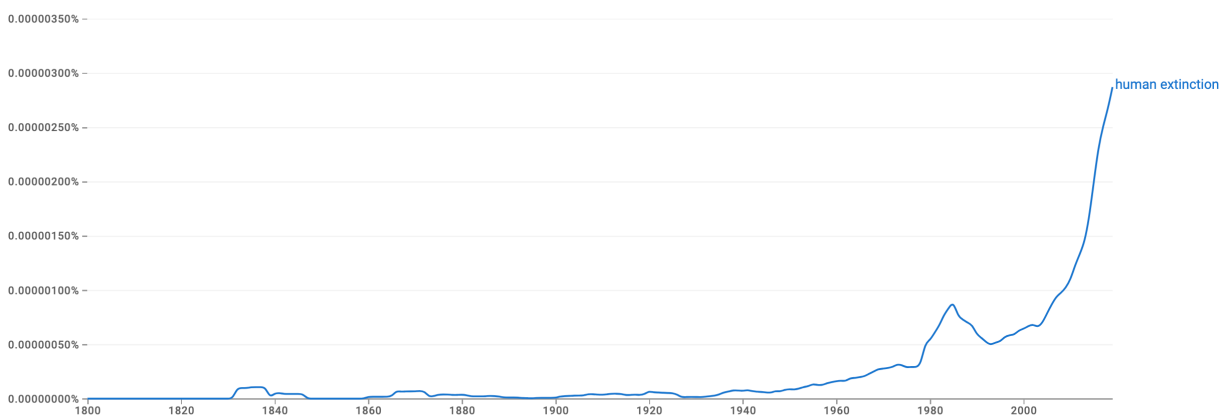


On the Extinction of Humanity

Abstract: The topic of human extinction is of growing importance and urgency. However, much of the discussion surrounding this topic is muddled by the fact that “human extinction” is highly polysemous: it can be defined in many different ways, and indeed different conceptions of human extinction can carry quite unique implications for how one assesses the ethical and evaluative implications of our collective disappearance. There are, furthermore, several additional distinctions that (a) are crucial for such assessments, and (b) no philosophers have yet made explicit in the literature—e.g., the distinction between the process or event of Going Extinct, and the state or condition of Being Extinct. This paper outlines a comprehensive framework for thinking about the ethical and evaluative aspects of human extinction, which may serve as a useful foundation for future research on the topic.

1. Introduction

The idea that humanity could someday disappear entirely from the planet can be traced back to the ancient Greeks, though it wasn’t until the 1950s that *human extinction* gained significant attention from intellectuals and the general public. A Google Ngram Viewer result for the keywords “human extinction” shows that the frequency of the term began to increase in the 1950s, underwent a significant spike in the 1980s, declined in the 1990s after the Cold War ended, and has increased more or less exponentially since the late 1990s (see figure 1). This is consistent with claims from numerous scholars that the probability of human extinction in the 21st century is unprecedentedly high—perhaps several orders of magnitude higher than at any period in our roughly 300,000 year history, with the one possible exception of the period following the Toba catastrophe circa 75,000 years ago, when the human population is hypothesized to have declined to ~10,000 individuals due to a volcanic winter caused by the Toba supereruption (Gibbons 1993). Nick Bostrom, for example, estimates a 20% chance of extinction before 2100, while a 2008 survey conducted by Bostrom’s Future of Humanity Institute put the probability at 19% (Bostrom 2005; Sandberg and Bostrom 2008). Along similar lines, Richard Posner declared in 2004 that “human extinction is becoming a feasible scientific project,” and that the near-term



risk of our extinction is “significant,” while the Doomsday Clock, maintained by the *Bulletin of the Atomic Scientists*, is currently set to a mere 90 seconds before midnight, or doom—the closest it has ever been since its creation in 1948 (Mecklin 2024). Surveys of the public mirror these anxieties. One reports that 39% of Americans believe that there is at least a 50% chance of climate change causing our extinction, while another 55% are “very” or “somewhat worried” that advanced AI will precipitate our collective demise (Leiserowitz et al. 2017; MUP 2023).¹

Consequently, public and academic discussions of human extinction—its potential etiology, its probability, and its ethical and evaluative implications—are gaining momentum, as indicated by the aforementioned Google Ngram Viewer results. There is, however, a major problem with these discussions: few commentators are clear about what “human extinction” could or should mean, and most people who use the term “human extinction” appear to be unaware that it is highly polysemous: both “human” and “extinction” could be defined in many different—and legitimate—ways, thus rendering “human extinction” doubly ambiguous.² This paper will focus on the various ways that “human extinction” can be defined with respect to what I call “Existential Ethics,” or the study of the ethical and evaluative implications of human extinction.³ I will show how certain ethical theories or normative frameworks may see one type of “human extinction” as being extremely bad or wrong, while simultaneously identifying other types of human extinction as neutral or even desirable. Other prominent positions within Existential Ethics are neutral about every type of “human extinction” except *one*, though this is not obvious or clear from the scholarly literature on such positions. I will then introduce some additional distinctions that are, I contend, crucial for making sense of the core questions of Existential Ethics, after which I will identify three major positions within the field, namely, further-loss views, equivalence views, and pro-extinctionist views. My aim is for this paper to lay the theoretical groundwork for future research on this increasingly important topic.

2. Disambiguating “Human” in “Human Extinction”

Consider that some of the philosophers who are most vocal about the importance of avoiding human extinction also endorse ethical positions that are either neutral about, or posi-

¹ My own view is that human extinction is extremely unlikely within the near future. I agree with Bruce Tonn’s claim that “the probability of human extinction is probably fairly low, maybe one chance in tens of millions to tens of billions, given humans’ abilities to adapt and survive” (Tonn 2009b). That said, I would contend, for reasons not elaborated here, that Martin Rees’ 2003 estimate that *civilization* has a 50/50 chance of collapsing this century is overly optimistic (Rees 2003). It is very difficult to imagine how our “civilization” can survive the profusion of unprecedented threats that will confront over the coming decades. But civilizational collapse need not entail human extinction.

² An exception is Elizabeth Finneron-Burns’ insightful 2024 article, which covers some of the same territory as my book chapter in *Redacted*. The present article elaborates on and further develops ideas first outlined in my aforementioned book chapter (*Redacted*).

³ The origins of Existential Ethics within the West date back to the 18th century (at least to Montesquieu’s 1721 *Persian Letters*), though the core questions of the field weren’t seriously addressed by more than a few people until the latter 19th century (within the tradition of German pessimism). The field underwent a growth spurt 100 years later, during the postwar era, following the Castle Bravo debacle in 1954, which convinced a large number of leading scientists and philosophers—for the first time—that human self-annihilation had become feasible. The most prominent contemporary theory within Existential Ethics is, arguably, longtermism, though this constitutes just one of many possible positions that one could hold with respect to the topic.

tively advocate, bringing about human extinction. Some transhumanists and longtermists, for example, would not object to the disappearance of “humanity” under certain circumstances, and indeed some actively call for the elimination of our species in the near future (see Thomas 2024). At the same time, they argue that avoiding “human extinction” should be one of our top global priorities this century. What is going on here? Drawing from my initiate discussion of this topic in [Redacted] (2023), let’s begin with a look at the different ways that “human” can be defined, and then turn to various “extinction” scenarios that could be instantiated given these different definitions of “human.”

The most obvious definition of “human” or “humanity” equates it with our biological species, *Homo sapiens*. The “extinction of humanity” would thus intuitively denote (something like) “the complete disappearance of the species *Homo sapiens*.” Let’s call this the *Narrow Definition*. However, many contemporary futurists prefer a more capacious definition. For example, Jason Matheny uses “‘humanity’ and ‘humans’ to mean our species and/or its descendants” (Matheny 2007). What does “descendants” mean? There are at least two possibilities: first, whatever beings come after us that are related to us in the right genealogical way. These beings might be biological, cyborgish, or wholly artificial, but so long as there is some kind of spatiotemporal continuity connecting us and them, they would constitute our descendants, the same way that we constitute the descendants of *Homo habilis*. Second, “descendants” could refer to whatever beings come after us that are related to us in the right causal way. This is more general than the first interpretation, as it makes room for, e.g., populations of intelligent machines that (a) take the place of *Homo sapiens*, and (b) we don’t evolve *into*, but rather create as separate entities from us. In the first case, an evolutionary lineage is preserved; in the second, one might say that a new lineage is created alongside us (and which could perhaps replace our lineage). Given that Matheny is writing within the transhumanist/longtermist tradition of “existential risk studies,” my guess it that Matheny meant to include both possibilities in his understanding of “descendants.” Hence, “human” refers to *Homo sapiens* and whatever future populations there may be that are related to us in the right genealogical or causal ways.

Other transhumanists and longtermists stipulate definitions of “humanity” that include an extra condition: that our descendants must possess one or more properties, in addition to those specified above, for them to count as “human.” For example, the longtermist Nick Beckstead writes that “by ‘humanity’ and ‘our descendants’ I don’t just mean the species [*Homo*] *sapiens*. I mean to include any valuable successors we might have,” which he later describes as “sentient beings that matter” in a moral sense (Beckstead 2013). Along similar lines, the longtermists Hilary Greaves and William MacAskill, who are sympathetic with transhumanism, report that “we will use ‘human’ to refer both to *Homo sapiens* and to whatever descendants with at least comparable moral status we may have, even if those descendants are a different species, and even if they are non-biological” (Greaves and MacAskill 2021). On these definitions, if there exist future populations of beings that are related to us in the right genealogical or causal ways but are not sentient or “valuable,” do not “matter” in a moral sense, or lack “at least comparable moral status,” then these future beings will not count as “human.” Let’s call the definitions of Beckstead (2013) and Greaves and MacAskill (2021), which build on Matheny’s definition, *Broad Definitions*.

An immediate conceptual, ontological, and terminological problem arises: many transhumanists and longtermists refer to future populations of beings that differ significantly from *Homo sapiens* by virtue of being “radically enhanced” in one or more ways as “posthumans” (see Bostrom 2008). In other words, if *Homo sapiens* were to use enhancement technologies to radically alter their cognition, healthspan, and so on, then the resulting beings would belong to a new taxonomic category called “posthumanity.” Yet so long as these future beings were to instantiate the properties of being our descendants with at least comparable moral status, they would also—by stipulative definition—constitute instances of “humanity.” Hence, certain future beings could simultaneously be “human” and “posthuman,” which appears to be contradictory. This conceptual/ontological/terminological confusion has some very substantial consequences: consider two groups, the humanists and the transhumanists. Let’s say that humanists are those who wish to preserve our species, *Homo sapiens*, into the foreseeable future; let’s say that transhumanists are those who wish to create one or more new posthuman species within the relatively near future. If asked, both would affirm that “avoiding human extinction is very important.” However, the humanists may understand “human” as denoting *Homo sapiens*, whereas the transhumanists would understand it as subsuming *Homo sapiens* and whatever posthuman beings we create or become. Note here that it may be possible to realize the transhumanist project of creating posthumanity without completely eliminating *Homo sapiens*: our species could continue to exist alongside the new posthuman species that we create or become.⁴ Some transhumanists refer to members of *Homo sapiens* that persist into the posthuman era as “legacy humans” (Goertzel 2010). Most transhumanists, though, would not object to *Homo sapiens* disappearing entirely (rather than living alongside their posthuman progeny), so long as *Homo sapiens* were replaced by a “superior” species of posthuman beings. Some transhumanists argue that we ought to completely replace *Homo sapiens* with a new species of posthuman. The transhumanist Steve Fuller, for example, has been described as “advocat[ing] an economics of death, whereby unaugmented humans (humanity1.0) may be sacrificed for the project of creating a superior successor species,” which he calls “humanity 2.0” (Thomas 2022). Similarly, Derek Shiller argues that “if it is within our power to provide a significantly better world for future generations at a comparatively small cost to ourselves, we have a strong moral reason to do so. One way of providing a significantly better world may involve replacing our species with something better” (Shiller 2017).

It follows that, from the humanists’ perspective, many transhumanists either either *okay* with human extinction occurring or positively *endorse* human extinction under such conditions. On the humanist view, such transhumanists accept a kind of *pro-extinctionism* (see below). In contrast, from the transhumanists’ perspective, *Homo sapiens* could disappear entirely and forever *without* human extinction having occurred. Indeed, on the Broad Definition, preventing human extinction does not entail that *Homo sapiens* must survive, except insofar as the survival of *Homo sapiens* is necessary to create future beings that satisfy the relevant conditions of “humanity.” This point is crucial because there might appear to be, at first glance, full agreement between humanists and transhumanists about the importance of avoiding human extinction, when in fact these views may be in direct conflict or, at the very least, in tension with each other.

⁴ By which I mean that some portion of *Homo sapiens* could become posthuman while the remaining members of our species would persist as they are.

To make this more concrete, consider a recent survey that reports “that most people find human extinction bad” and identify “the prevention of human extinction [to be] a key priority” (Coleman et al., unpublished). This study was coauthored by two scholars previously based at the Global Priorities Institute, a longtermist/transhumanist organization at the University of Oxford, and hence it seems probable that these scholars are tacitly adopting the Broad Definition. Yet most people, I would conjecture, will intuitively adopt the Narrow Definition. Hence, the survey’s results may give the impression that most members of the general public concur with the longtermists and transhumanists about preventing human extinction being a “key priority,” when in fact many people would probably be consternated and alarmed to discover that *Homo sapiens* might not have a place in the longtermist-transhumanist vision of the future. (As Toby Ord (2020) writes, “forever preserving humanity as it is now may ... squander our legacy, relinquishing a greater part of our potential” in the universe.) This is one reason that disambiguating the word “humanity” is so important.

3. Disambiguating “Extinction” in “Human Extinction”

There are also several distinct types of extinction scenarios that are directly relevant to assessments of the ethical and evaluative implications of our extinction, in addition to the different possible definitions of “humanity.” Some of these scenarios apply to both the Narrow and Broad definitions, while others are unique to one or the other definition. Let’s begin with a Minimal Definition of “extinction,” and then examine the various scenarios that build on this Minimal Definition with respect to the different ways that “humanity” could be defined.

Minimal Definition: a type of thing S has gone extinct if and only if there were tokens of S at some time T1, but, at some later time T2, no such tokens exist.

This definition is straightforward and intuitive, and could apply to a wide range of things, including languages, cultures, and biological species. It also corresponds to the standard definitions of “extinction” that one finds in dictionaries and science textbooks, as when Merriam-Webster defines “extinction” as “the condition or fact of being extinct,” and “extinct” as “no longer existing” (Merriam-Webster 2021). Similarly, Julien Delord writes that “most biologists accept the following basic definition: ‘The end, the loss of existence, the disappearance of a species or the ending of a reproductive lineage’” (Delord 2007). This leaves open the possibility of this disappearance being *temporary*, and indeed one of the goals of the emerging field of Resurrection Biology is to bring back into existence species that are no longer instantiated in the world. Let’s call this *demographic extinction*, noting that, while it may be possible for humanity (however defined) to disappear in the future and later on be resurrected from the grave, the circumstances

in which this might happen are highly speculative and very improbable. (Demographic extinction is primarily of historical interest.⁵)

A stronger conception of human extinction would add a condition of permanence: the extinction of humanity would occur, on this conception, if and only if humanity were to disappear entirely *and forever*. Let's call this *terminal extinction*. This could happen on both the Narrow and Broad definitions of "humanity." In the former case, it would entail that there are no more instances of *Homo sapiens* in the world, and that this state of affairs never reverses; in the latter case, it would entail that there are no more instances of the class of beings that includes *Homo sapiens* and whatever descendants (as defined) that we might have with comparable moral status, whether biological or artificial, and that this state of affairs never reverses. Terminal extinction is complete and permanent, though the Broad Definition implies that *Homo sapiens* itself can disappear entirely and forever without *humanity* having disappeared entirely and forever. Indeed, this is precisely where humanists and transhumanists will diverge: the former opposes *Homo sapiens* going terminally extinct (on the Narrow Definition), whereas the latter opposes this *unless* we have successors with the right properties.

This gets at a second category of extinction scenario, which I will call *final extinction*. Assume for a moment the Narrow Definition of "humanity." Final extinction would occur if and only if the following conditions were met: *Homo sapiens* disappears entirely and forever, *and* we do not leave behind any successors (the nature of these successors being left unspecified, for reasons explained below). The significance of this definition, in contrast to terminal extinction, is that what happens *after* our particular species has disappeared could make a big difference with respect to how one assesses the rightness/wrongness or goodness/badness of the outcome. There are many ethical perspectives and value theories that would judge the terminal extinction of *Homo sapiens* to be bad *only* if this were to coincide with, or entail, there being no future descendants of ours. Otherwise, it would not—which is just to say that these ethical perspectives and value theories do not identify the terminal extinction of *Homo sapiens* itself as intrinsically bad, only instrumentally bad; the scenario they identify as unconditionally bad is final human extinction.

To illustrate, consider Samuel Scheffler's (2018) argument (simplified here) that the prospect of near-term "human extinction" would cause many people to collapse into despair, finding many of the activities that once gave value to their lives empty and purposeless. One reason is that an important source of value in our lives derives from meliorative *transgenerational* projects that, as such, can only be brought to fruition if there are future people who carry on these projects. The question thus becomes: must these future people be members of *Homo sapiens*? The answer appears to be "no": one can imagine posthuman beings that (a) are, as such, not members of *Homo sapiens*, (b) entirely replace *Homo sapiens*, and (c) carry on such projects. (Perhaps these future "people" would be intelligent, conscious machines that continue the work of science, further develop and appreciate the arts, and so on.) Hence, using the Narrow Defini-

⁵ For example, several Presocratic philosophers, such as Xenophanes, proposed cosmological models in which the cosmos cycles through different stages. During one of these stages, all human beings on Earth perish. At a later stage, though, we will always reemerge. In other words, Xenophanes believed that our species will someday undergo demographic extinction, but not *terminal* extinction (a concept introduced just below): our complete disappearance from the universe is always a temporary state of affairs.

tion, does Scheffler's argument imply that we must avoid terminal extinction? Yes, but only instrumentally: if *Homo sapiens* were to undergo terminal extinction, this would not in itself undermine the various transgenerational projects that enable us to have "value-laden" lives today. What we must avoid is final extinction, which *would* entail the destruction of these projects. Scheffler's discussion of why avoiding "human extinction" matters is thus ambiguous between these two quite distinct scenarios: though he doesn't say it, his arguments target final rather than terminal extinction, on the Narrow Definition.

Another example comes from totalist utilitarianism, which is intimately linked to longtermism (as longtermism, even in its "moderate" forms, assumes the axiology of totalism; see MacAskill 2022). If our sole moral obligation is to maximize welfare, and if a certain kind of posthuman species would be better able to maximize welfare than *Homo sapiens*, then the totalist utilitarian should positively hope that *Homo sapiens* goes out of existence in the future—*so long as* this coincides with the emergence of a new posthuman species that is better able to maximize welfare. In other words, the terminal extinction of *Homo sapiens* would not be intrinsically bad, though the final extinction of *Homo sapiens* (very much) could be. The only condition in which terminal extinction would be very bad is if it were to simultaneously instantiate final extinction. To make the relationship between these two scenarios more explicit, then: final extinction entails terminal extinction, but terminal extinction need not entail final extinction. Similar points could be made about longtermism and transhumanism, as well as various other "further-loss views," explained below.

One last example is worth mentioning: consider the position defended by David Benatar (2006), according to which we should bring about our extinction as soon as possible. The reason he gives is that (a) birth is always a net harm, and (b) life is overflowing with misery, and though many of our lives are not so bad as to warrant no longer continuing them, they are sufficiently bad as to warrant not having started them. (a) Does not imply that humanity should go extinct, because there could be future technologies that enable people to live forever, and hence for humanity to indefinitely persist even in the absence of procreation. Let us, therefore, focus on (b). If the claim is that future beings, whether posthuman or instances of *Homo sapiens*, will have lives that are very bad (and I believe that this is much more plausible than one might initially suspect), then which type of extinction should Benatar advocate for: terminal or final? Clearly, Benatar should advocate for final human extinction, as this would constitute the only way of guaranteeing that large amounts of human suffering that would otherwise exist in the future will not exist. The same conclusion applies to other philosophers in the pessimist tradition, such as Philipp Mainländer, Eduard von Hartmann, and Peter Wessel Zapffe, as well as to negative utilitarians and radical environmentalists who endorse human extinction for ecological reasons (often based on biocentric, biospherical-egalitarian, or ecocentric theories of value). Terminal extinction itself would not guarantee the elimination of future suffering—our sole moral obligation according to negative utilitarianism—nor is it sufficient to ensure that the obliteration of the biosphere stops. If, for example, *Homo sapiens* disappears but we create or become a successor species that carries on our environmentally destructive activities, the problem that "extinction" is supposed to solve will persist. Hence, the target of all these perspectives is, by implication, final rather than terminal extinction.

I hope to have shown at this point that distinguishing between these two extinction scenarios is very important. However, if one begins with the Broad Definition of “humanity,” this distinction collapses. Since final extinction references what comes after “humanity” has disappeared, but if “humanity” is defined, in part, as *Homo sapiens* and whatever might come after us, then terminal extinction *just is* final extinction. If “humanity” on the Broad Definition disappears entirely and forever—full stop—then there are no more instances of *Homo sapiens* or descendants of *Homo sapiens*. On the Broad Definition, then, what matters to Scheffler, totalist utilitarians, longtermists, Benatar, negative utilitarians, and some radical environmentalists is terminal extinction. For the first few in that list, what matters is avoiding terminal extinction, given the Broad Definition, while for the last few, what matters is bringing about terminal extinction. However, for humanists (as I am here using that term), talk of “terminal extinction” using the Broad Definition is a problematic framing, since what humanists care about is our particular species, *Homo sapiens*, persisting into the long-term future.

This is why, once again, disambiguating the term “humanity” is so important, and why, once this term is disambiguated, specifying which type of “extinction” one is concerned about is crucial. On the Narrow Definition, humanists will oppose terminal extinction, independent of whether it coincides with final extinction, while longtermists (for example) will be instrumentally invested in avoiding terminal extinction because what they ultimately care about is avoiding final extinction. On the Broad Definition, humanists will complain that the framing is all wrong, while longtermists (again, for example) will say that terminal extinction—the complete and permanent disappearance of “humanity,” on this more capacious definition—is what we must avoid at any cost.⁶

There are two additional extinction scenarios, one of which uniquely pertains to the Narrow Definition, while the other mostly pertains to the Broad Definition. The first is what could be called *phyletic extinction*. This would occur if and only if *Homo sapiens* were to evolve into one or more posthuman species through a process that preserves, as it were, the spatiotemporal continuity of our evolutionary lineage. As with every other type of extinction here discussed, this satisfies the Minimal Definition of “extinction,” in the following way: at T1, there exist instances of *Homo sapiens* in the world but, at some later time T2, our species evolves into a successor population that is sufficiently different for it to warrant classification as a novel species; no more instances of *Homo sapiens* remain. There are a couple of immediate points to make about this: first, phyletic extinction foregrounds the concept of *species*, which is hotly debated among philosophers of biology and evolutionary biologists. There is no need to settle this debate here: suffice it to say that, however one defines “species,” phyletic extinction would happen if future populations of our descendants satisfy one’s preferred definition of “species,” such that these future beings no longer count as instances of *Homo sapiens*. That is the crux of phyletic extinction. Second, the case of *Homo sapiens* is completely unique within the Animal Kingdom, as we appear to be on the verge of developing technologies that could enable us to radically alter our phenotypic traits. In the long run, phyletic extinction is inevitable due to evolutionary mechanisms like natural selection, random mutation, genetic drift, recombination, and so on. However, in the

⁶ I borrow the phrase “at any cost” from Bostrom, who writes that, from a transhumanist (and longtermist) perspective, “there is one kind of catastrophe that must be avoided at any cost: *Existential risk*,” where an “existential risk” could take the form of “human extinction” (Bostrom 2005).

short run, we may reengineer ourselves via some process of cyborgization, where the limit of this process would be the complete replacement of our biological substrate with an artificial substrate (as in the case of “mind uploading”). Put differently, there are, in the case of *Homo sapiens*, two general possibilities for how phyletic extinction could occur: one natural, and the other anthropogenic.

We now have three different conceptions of human “extinction.” When someone asks the question—for example—“Would human extinction be bad?,” one should respond with two clarificatory questions: “What do you mean by ‘human’?” and “Which type of ‘extinction’ are you referring to?” Some people who prefer the Narrow Definition may claim that avoiding final human extinction is what ultimately matters, while adding that phyletic extinction, if the result is a “superior” new species of posthumans, would be positively desirable. Some transhumanists would champion this position. Others may say that avoiding terminal extinction is what ultimately matters, and that since phyletic extinction would—according to our definition above—entail that *Homo sapiens* no longer exists (a situation that may be permanent), we should oppose phyletic extinction. Someone who holds this view might follow Johann Frick (2017) in arguing that *Homo sapiens* has final value because of its uniqueness, and hence that the loss of this value would render the universe more impoverished. Perhaps the posthuman species that takes our place would also be unique, but not unique in the same way, and hence something would still be lost even if our successors were finally valuable. This has the curious implication that those who hold this view should oppose phyletic extinction via natural processes, too, perhaps by using advanced genetic engineering technologies to “fix” our genotypes within some specified range of variability, thereby preserving *Homo sapiens* into the distant future. Still others might introduce a timescale constraint to their evaluations of human extinction, arguing that we should avoid terminal extinction within the foreseeable future, but if phyletic extinction were to unfold gradually enough—over hundreds of thousands or millions of years—it would not be bad. (My guess is that many people would intuitively endorse this particular view: if there were no more instances of *Homo sapiens* in the future, that would not be bad, or the badness would be greatly mitigated, if the reason this happened is that we incrementally evolved into a new species tens of thousands or millions of years from now.)

Notice that phyletic extinction is not applicable to the Broad Definitions delineated above (which is, again, why these definitions matter). If “humanity” means *Homo sapiens* plus our descendants, then it cannot be the case that “humanity” disappears by evolving into a successor species, as any such successor species would also count as “humanity.” However, there is another type of extinction that primarily concerns the Broad Definition. Consider a scenario in which we have have descendants, but these descendants come to lack the capacity for conscious experience. Assuming that consciousness is a necessary condition for something to have any moral status, if one defines “humanity” as “our species or whatever successors we may have, given that they possess at least comparable moral status to us,” these future beings would not count as instances of “humanity.” And if these future beings do not count as “human” on the definitions proposed by Beckstead (2013) and Greaves and MacAskill (2021), then the Minimal Definition of “extinction” would be satisfied. Hence, given these Broad Definitions, there would be no more humans in the world, even if there existed future beings that are genealogically or casually

related to us in the relevant ways. Let's call this *normative extinction*, since it concerns (primarily) the normatively substantial Broad Definitions.⁷

Avoiding normative extinction is of great importance from certain ethical and evaluative perspectives. Take totalist utilitarianism, for example. One of the motivations behind Beckstead's and Greaves and MacAskill's normative definitions of "humanity" is the axiological claim that the world is better the more value that it contains. A world without humans would contain much less value, while a world full of very larger numbers of humans, so long as they have (at least) net-positive lives, would contain much more. Hence, it matters not only that we have descendants, but that these descendants count as "human" (even more, it matters that these future "humans" are also posthumans, as posthumans may be capable of bringing far more value into the world than members of *Homo sapiens* can.) Assuming their Broad Definition, the loss of the value that might occur if we have descendants that aren't "human" could be comparable to the loss of value the would occur if we do not have any descendants at all. From this perspective, therefore, normative extinction may be equivalent in badness/wrongness as terminal extinction (again, on the Broad Definition, which is the same as final extinction on the Narrow Definition).

It follows that, for totalist utilitarians, the most relevant definitions of "humanity" are the normatively substantial Broad Definitions, and the two types of extinction that would guarantee a failure to maximize value are terminal and normative extinction. The complete and permanent disappearance of *Homo sapiens* is not intrinsically bad—only instrumentally bad, if it occurs in such a way that it prevents us from having successors who matter morally—while the phyletic extinction of *Homo sapiens* may be positively desirable, if it means that a new species of posthumans comes into being that is better able to generate and spread welfare (perhaps by virtue of these posthumans being digital, which would enable them to journey beyond our solar system—something that is probably impossible for biological humans; see [redacted]). The same could be said about other accounts of human extinction, such as that defended by Hans Jonas (1979). On Jonas' view, what matters is the perpetuation of what he calls the "moral universe"

⁷ One commenter on this paper wrote that "it is not clear that we then need different concepts of 'extinction.' We only get different versions of [extinction] when we consider different versions of [humanity]. Can't we just talk about different notions of 'human' and consider then which one of these notions a concern about 'extinction' applies to?" As I try to show in this article, disambiguating the term "extinction" is no less important than clarifying the term "humanity," as there are multiple types of extinction, each carrying its own unique ethical and evaluative implications, associated with different definitions of "humanity." Consider the case of normative extinction. In his 2013 paper on existential risks, Bostrom identifies "human extinction" as one type of existential catastrophe. But what Bostrom means by "human extinction" is "terminal human extinction on the Broad Definition of 'humanity.'" Later in his article, he imagines a future in which we have successors that lack conscious experiences—they are philosophical zombies who nonetheless carry on the project of civilization (science, the arts, etc.). This would qualify as a *different* type of existential catastrophe that he calls "flawed realization." Yet, on the Broad Definitions that are popular among transhumanists and longtermists, this would actually constitute *normative extinction*. It would constitute normative extinction because, given a Broad Definition of "humanity" and the minimal definition of "extinction," humanity would no longer exist in this scenario. This is one of many reasons for, I would argue, being more precise about our terms and concepts. When Bostrom talks about this type of "flawed realization," he is actually talking about human extinction in the "normative" sense. Without a clear understanding of the notions of both *humanity* and *extinction*, one cannot make sense of the claim—which I elaborate in Redacted—that many transhumanists, longtermists, and especially "accelerationists" are actually "pro-extinctionists," despite being among the loudest voices advocating for the *avoidance* of "human extinction." Such individuals are in favor of, or are at least indifferent to, the near-term terminal extinction of our species, while simultaneously strongly opposing the final extinction of humanity, on the Narrow Definition.

within the physical universe. The moral universe is made possible by the existence of moral agents: beings with the capacity for moral responsibility, of which we are the only known instance. Although readers of Jonas' work may assume that his view requires the continuation of *Homo sapiens*, this is not the case: the only two types of human extinction that are sufficient for eliminating the moral universe are terminal and normative extinction—or, if one adopts the Narrow Definition, final rather than terminal extinction, as the continuation of *Homo sapiens* is itself not a necessary condition for the moral universe to persist. See figure 2 for a diagrammatic representation of these distinctions.

Narrow Definition Humanity = <i>Homo sapiens</i>	Broad Definition Humanity = <i>Homo sapiens</i> or whatever descendants we might have with the right moral status
Terminal Extinction Humanity disappears entirely and forever	Terminal Extinction Humanity disappears entirely and forever
Final Extinction Humanity disappears entirely and forever, without leaving behind any successors	Normative Extinction Our species has descendants, but these descendants lack something that is normatively required for them to count as "human"
Phyletic Extinction Humanity evolves into one or more posthuman species	

This brings us to a final set of distinctions that are crucial for navigating the labyrinth of Existential Ethics, though before turning to this, let's note one last type of human extinction scenario that could apply to both the Narrow and Broad definitions: *premature extinction*. The term "premature extinction" was initially used by ecologists to describe the loss of a nonhuman species due to anthropogenic causes—i.e., before the species would have otherwise "naturally" gone extinct. So far as I am aware, the first futurist to use this term in the context of Existential Ethics (and hence in the context of *human* extinction) was Bruce Tonn in 2009; it was then foregrounded and popularized by Bostrom and Beckstead in two separate publications, both from 2013 (Tonn 2009a; Bostrom 2013; Beckstead 2013). Premature extinction introduces the idea that the *timing* of human extinction, however defined, matters morally. It is not an alternative to terminal, final, and normative extinction, but rather describes a particular *way* that these types of extinction could happen. The claim would be that if, say, the terminal extinction of humanity, on

the Broad Definition, were to happen *after* attaining some desired goal or completing some valued project, it may still be very bad, but it would be *less bad* than if this were to happen *before* attaining that goal or completing that project. *When* our extinction happens is important, those who endorse premature extinction would argue. For longtermists, premature extinction might occur if we undergo terminal/final extinction (depending on the definition of “humanity”) prior to realizing a large fraction of “our longterm potential” in the universe. Bostrom defines it as when humanity—by which he means “Earth-originating intelligent life,” an even more capacious definition than Beckstead’s or Greaves and MacAskill’s—“goes extinct ... before reaching technological maturity,” i.e., a state in which we achieve “capabilities affording a level of economic productivity and control over nature close to the maximum that could feasibly be achieved” (Bostrom 2013).

Others would define premature extinction as something like terminal, final, or normative extinction that happens before “humanity” finishes certain “business,” such as the business of devising a complete scientific theory of the universe. In fact, the first reference to the idea of premature extinction that I am aware of comes from a 1978 paper by Jonathan Bennett, in which he suggests that we have a “prima facie obligation to ensure that important business is not left unfinished.” Some have called this the “unfinished business argument” for ensuring the survival of humanity, where the most natural interpretation of “humanity” is along the lines of the Broad Definition, given that finishing much of our current business (e.g., science) does not obviously require *Homo sapiens* itself to exist; such projects could instead be carried on by some suitable posthuman successor. In sum, the idea of premature extinction may be invoked whenever one accepts a vision of our future that is both normative and teleological, which is to say: a vision that identifies a future goal, or *telos*, as something that we ought to strive for. For the purposes of this paper, I will bracket the idea of premature extinction in what follows.

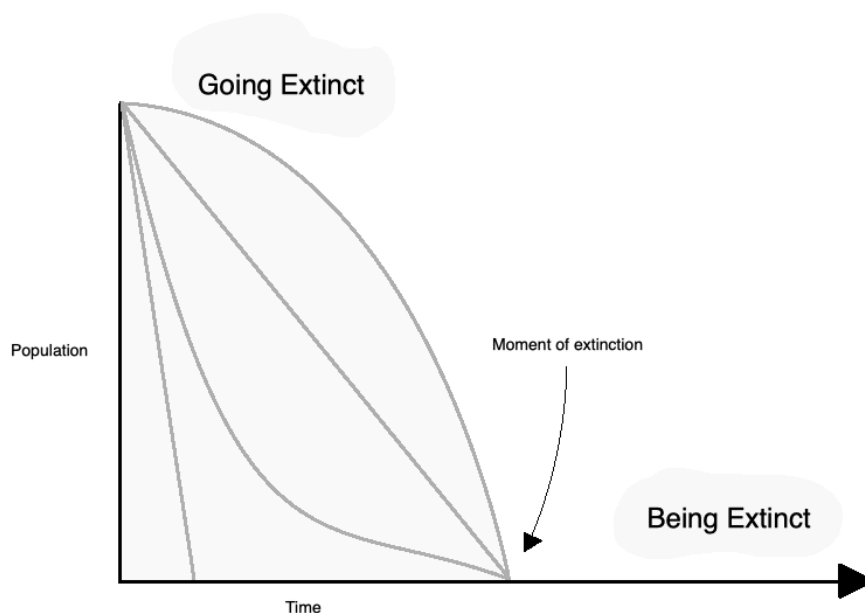
4. Going Extinct Versus Being Extinct

We have now established that the ambiguities of “human” and “extinction” have important implications for the ethical and evaluative assessment of human extinction. To date, no philosophers have made these ambiguities explicit or noted their critical relevance to Existential Ethics. This section turns to another set of distinctions that have been similarly neglected, and which are no less crucial for understanding the different positions that one could take within Existential Ethics. The distinction is between the process or event of Going Extinct, and the state or condition of Being Extinct. This could apply to all of the types of extinction discussed above; hence, there is Going terminally Extinct and Being terminally Extinct; Going normatively Extinct and Being normatively Extinct, and so on. As we will see in the next section, some ethical and value theories claim that the badness or wrongness of human extinction comes down entirely to the details of Going Extinct, whereas others identify Being Extinct as a source of badness/wrongness as well. Still other theories see most ways of Going Extinct as bad or wrong, while simultaneously claiming that Being Extinct would in some sense be better than Being Extant, i.e., continuing to exist.

This gestures at a second set of distinctions, which specifically concern the phenomenon of Going Extinct: it may also matter morally *how exactly* this process or event unfolds. There

are, of course, many different ways that humanity could go extinct, however defined. Consider the final extinction of *Homo sapiens*: this could result from a large asteroid striking Earth and inducing an impact winter. It could result from an engineered pandemic involving designer pathogens synthesized in a biohacker laboratory and intentionally released into the world. It could result from an involuntary infertility scenario in which a sufficient number of people around the world are unable to procreate such that the human population dwindles to zero. Or it could result from everyone around the world voluntarily deciding not to have children (perhaps because they read the works of von Hartmann or Zapffe). The *etiology* of Going Extinct is thus important: some existential ethicists would say that final human extinction caused by an engineered pandemic would constitute a terrible moral crime, but that there wouldn't be anything bad or wrong if this were entirely voluntary. Others would vociferously disagree, claiming that the details of Going Extinct are only part of the picture—perhaps a relatively small part—of why our disappearance would be bad or wrong. With respect to natural versus anthropogenic causes of extinction, only the latter is relevant to ethics, as a natural cause of extinction may be judged to be bad, but there is nothing morally wrong with asteroids hitting Earth or viruses that evolve in nature killing everyone on Earth.

Another property of Going Extinct that may matter morally is its *temporality*. For example, Hermann Vetter argued in 1968 that “if mankind were extinguished by a nuclear war, the real evil ... would be the way the extinction would take place: there would be so much terrible suffering for so many people before they die that this is a tremendous evil,” but “if mankind were completely extinguished in a millionth of a second without any suffering imposed on anybody, I should not consider this as an evil, but rather as the attainment of Nirvana” (Vetter 1968). On Vetter's account, *instantaneous* extinction would be good, while *drawn-out* extinction would be bad, if it were to cause lots of human suffering. Negative utilitarians would agree, as some pessimists who endorse human extinction (though this depends in part on whether one accepts an



anti-Epicurean view of death; if one accepts such a view, as I do, then instantaneous extinction could still cause tremendous harm even if no one experienced any physical or psychological suffering).

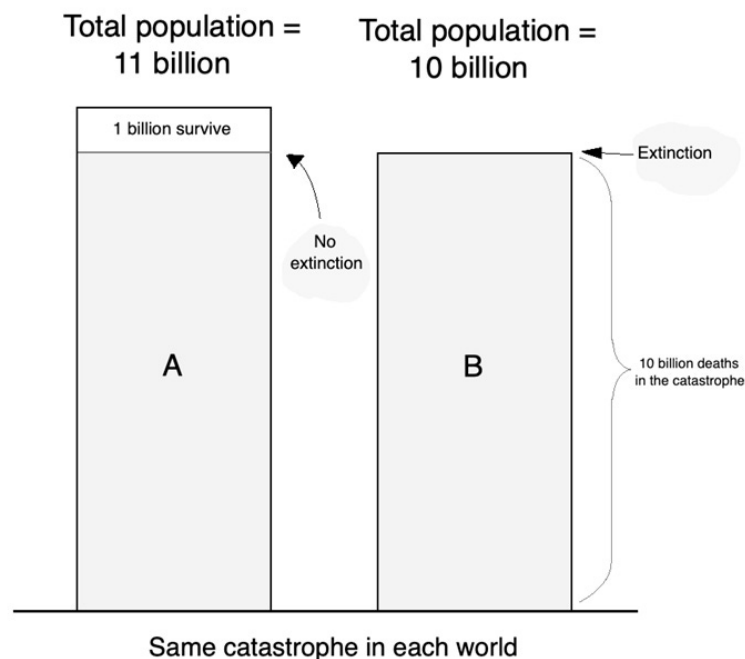
The key points of this short section are: assessing whether human extinction is right or wrong, good or bad, better or worse crucially depends on separating out the process or event of Going Extinct and the subsequent state or condition of Being Extinct. Furthermore, there are several important distinctions to make with respect to Going Extinct, such as whether it is natural or anthropogenic, voluntary or involuntary, and instantaneous or drawn-out. (See figure 3 for a visual illustration of different scenarios of final human extinction, assuming the Narrow Definition.) This second cluster of distinctions raises many new questions, such as “What exactly is the difference between ‘natural’ and ‘anthropogenic’ causes?” and “What does it mean for our extinction to be ‘voluntary’? Would extinction be voluntary if 98% of people endorsed it but 2% vehemently objected?” I will not pursue these interesting and important questions here due to space constraints.

5. Three Main Positions within Existential Ethics

Building on the previous section, we can identify three main positions within Existential Ethics. The first class of positions is what I call “equivalence views.” The hallmark of such views is the claim that the badness/wrongness of human extinction, however defined, is reducible entirely to the details of Going Extinct. If there is something bad or wrong about Going Extinct, then there is something bad or wrong about our extinction; if there is nothing bad or wrong about Going Extinct, then there is nothing bad or wrong about our extinction. That is the end of the story—there is nothing of moral significance about the state or condition of Being Extinct. It is irrelevant. Hence, if human extinction were to happen because an engineered pandemic sweeps across the globe and slowly kills everyone, then it would be very bad and/or very wrong. But if everyone around the world were to voluntarily decide not to produce children, resulting in the human population falling to zero over the course of ~120 years, this would not be bad or wrong. Person-affecting theories like Scanlonian contractualism and (total) utilitarianism of the sort championed by Jan Narveson (1967)—a kind of person-affecting or personalist utilitarianism, in contrast to the impersonalist utilitarianism of Henry Sidgwick—are examples of equivalence views. On at least some of these theories, the central claim is that Being Extinct is morally irrelevant because it would necessarily entail that there are no more humans in the world; and if there are no humans to be harmed by the absence of things such as happiness, satisfied desires, and so-called “ideal goods” like the sciences and the arts, then where is the wrong? Where is the badness? Although Going Extinct may involve great horrors and/or grave injustices, thus making our extinction very bad or wrong, Being Extinct is neither bad nor wrong.

An interesting implications of equivalence views is that there is no unique problem posed by our extinction. There is nothing to say about *extinction itself*. In other words, everything one might wish to say about human extinction can be said without any reference to “extinction” at all. If *Homo sapiens* undergoes final (or phyletic, or normative, etc.) extinction voluntarily, then nothing bad or wrong has happened; if *Homo sapiens* undergoes final extinction through some catastrophe, then that would be very bad because catastrophes are very bad. “Extinction” as a

result of a catastrophe is just the name we give to the limit of how bad such catastrophes could be: “extinction by catastrophe” simply conveys the fact that this has the highest body count possible. That may make it the worst type of catastrophe that could possibly occur, but not because it results in our *extinction*—rather, because it claims the maximum number of casualties. To illustrate, consider the following two worlds: in World A, there are 11 billion people; in World B, there are 10 billion people; see figure 4. An identical event happens in both worlds that kills exactly 10 billion people. At a high level of abstraction, we can ask: how many events happen in each of these worlds? In World A, a single event happens: the death of 10 billion people. In World B, two events happen: the death of 10 billion people and the extinction of humanity. The second question is thus: does this extra event in World B make any ethical or evaluative difference? If the event in both worlds is a global catastrophe, is the scenario of World B *worse* than that of World A? Or, if an omnicidal maniac named Joe kills 10 billion people in both worlds, does he do something *extra* wrong in World B?



Equivalence theorists, as we can call them, will claim that the badness of the events in both worlds—assuming, again, that these events are identical—is the exact same. The World B scenario is not worse than the World A scenario, but equivalent. Similarly, they will claim that Joe does not do something extra wrong in World B compared to World A; the wrongness in each is also equivalent.⁸ The fact that extinction happens in World B is not ethically or evaluatively relevant, precisely because the state or condition of Being Extinct is not relevant. Assessing the badness/wrongness of human extinction comes down entirely to the details of Going Extinct.

A second class of positions within Existential Ethics, which I call “further-loss views,” strongly disagree. They claim that Being Extinct can also be a source of badness/wrongness, and

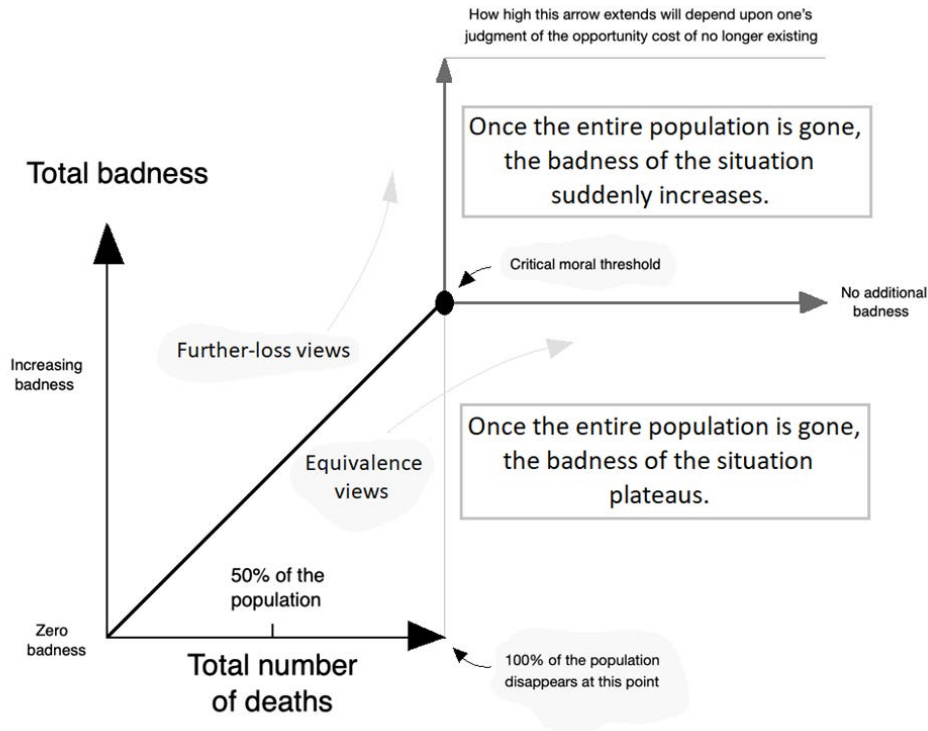
⁸ This points to the possibility that equivalence views could have both deontic and evaluative interpretations. I will not pursue this nuance here.

hence that assessing human extinction scenarios is a two-step process: first, one examines the details of Going Extinct. Does the process or event of Going Extinct cause physical or psychological suffering? Does the way in which this unfolds conflict with the morally relevant interests of those living at the time? And so on. Second, one lists various further-losses associated with the state or condition of Being Extinct that one takes to be ethically and/or evaluatively important. Totalist utilitarians, for example, would point to all the “lost” value that could have otherwise existed over the next millions, billions, and trillions of years if humanity had avoided (let’s say) final extinction. The same goes for longtermists: the “opportunity cost” of our species suddenly disappearing, say, two decades from now would be the enormous number of future generations that would never exist.

However, one does not need to be a totalist utilitarian or longtermist to accept a further-loss account of the badness/wrongness of human extinction. Earlier I referenced Hans Jonas’ claim that we have a moral obligation to ensure the perpetuation of the moral universe within the physical universe. This is a further-loss view because Jonas is claiming that the badness/wrongness of human extinction goes *above and beyond* whatever harms the process or event of Going Extinct might involve. The obliteration of the moral universe constitutes an ethically important further-loss that would make our (final or normative) extinction very bad *independent* of the details of Going Extinct. Or consider Scheffler’s assertion that one reason we should work to ensure our continued survival into the indefinite future is that many of the things we value in the world cannot exist without us. If humanity were to disappear, so would these things; that is a problem because, as Scheffler writes, “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?” (Scheffler 2007). The potential loss of these things if humanity goes extinct thus gives us reason to avoid this scenario. This, too, gestures at a further-loss position. Perhaps the first reference to a further-loss view in the Existential Ethics literature comes from Mary Shelley’s 1826 novel *The Last Man*, in which the main character, Lionel Verney, reflects on the tragedy of extinction, noting that our disappearance (a scenario of final extinction brought about by a worldwide pandemic) would mean not only that there are no humans in the future, but that there no longer exist valued things like knowledge, science, technology, poetry, philosophy, sculpture, painting, music, theater, laughter, and so on (Shelley 1826).

In the case of figure 4, further-loss theorists would argue that the scenario of World B is much worse than the scenario of World A, and that Joe does something extra wrong in World B compared to World A. *How much* worse the scenario of World B is will depend on the significance of these further-losses. Many further-loss theorists will argue that the additional losses or opportunity costs associated with Being Extinct are *far greater* than whatever harms Going Extinct might—or could possibly—involve. As Peter Singer, Beckstead, and Matthew Wage write:

One very bad thing about human extinction would be that billions of people would likely die painful deaths. But in our view, this is, by far, not the worst thing about human extinction. The worst thing about human extinction is that there would be no future generations (Singer et al. 2013).



This implies that human extinction is a qualitatively unique kind of tragedy, and hence that our extinction—by which most further-loss theorists will mean “final” extinction on the Narrow Definition, or “terminal” and “normative” on the Broad Definition—*does* constitute a unique moral problem. One expression of this idea comes from Derek Parfit’s thought experiment at the end of his 1984 book *Reasons and Persons*. He asks us to imagine three scenarios: “(1) Peace. (2) A nuclear war that kills 99% of the world’s existing population. (3) A nuclear war that kills 100%” (Parfit 1984). Which is the greater *difference in badness*: between (1) and (2), or between (2) and (3)? He claims that most people would say the first difference is the greatest, and indeed one recent survey of members of the general public suggests that Parfit’s hypothesis is correct (see below; Schubert 2019). Equivalence theorists will agree, arguing that the difference between (2) and (3) is just 1% of humanity perishing in a catastrophe—end of story. However, Parfit contends that the greater difference is (2) and (3), because the scenario of (3) precludes the realization (a) all future happiness (which could be substantial), and (b) the further development of ideal goods like science, the arts, and morality. Both (a) and (b) are further-losses that, in Parfit’s view, make human extinction qualitatively different from non-extinction scenarios in which most, but not all, of humanity perishes.

Another way to highlight the difference between equivalence and further-loss views is as follows: imagine a catastrophe that, over a period of 1 year, kills more and more people until humanity has undergone final extinction (or terminal extinction, on the Broad Definition); see figure 5. The equivalence and further-loss theorists might agree that as the number of casualties increases, so does the badness of the catastrophe; in the simplest case, twice as many deaths makes the catastrophe twice as bad. However, something important happens once the number of

casualties reaches the maximum: for equivalence theorists, the badness of the situation plateaus. Why? Because the moment of extinction marks the transition from Going Extinct to Being Extinct, and Being Extinct is ethically and evaluatively irrelevant. For further-loss theorists, the badness of the situation suddenly skyrockets. Why? Because the moment of extinction marks the point at which all future value is lost forever. How high the vertical line extends from the “critical moral threshold” of extinction depends on how great one judges the further-losses to be. If one accepts Bostrom’s estimate that there could be 10^{58} “happy” digital people in the future, one might extend the vertical line hundreds or thousands of feet above the diagram as shown in this article. Similarly, from my reading of Jonas I get the impression that Jonas considers the loss of the moral universe to be *very bad*, and hence he, too, may extend the vertical line much further than its current length in the figure. For many further-loss theorists, Being Extinct is not only ethically and evaluatively relevant, but the badness/wrongness associated with Being Extinct far exceeds the badness/wrongness of even the most terrible ways that Going Extinct might unfold.

An important implication of further-loss views is that human extinction—for simplicity, let’s focus on final human extinction, on the Narrow Definition—could be very bad or wrong *even if* there is nothing bad or wrong about Going Extinct. In figure 4, imagine that 10 billion people in each world voluntarily decide not to have children. Equivalence and further-loss theorists would (presumably) agree that there is nothing bad or wrong about this scenario in World A. However, they would vehemently disagree about the badness/wrongness of this scenario in World B, since the voluntary decision of 10 billion people not to procreate in World B would result in our extinction. As Sidgwick famously declared in a discussion of voluntary celibacy, there is nothing obviously wrong with any given individual choosing celibacy, yet “a *universal refusal* to propagate the human species would be the greatest of conceivable crimes from a [totalist or impersonalist] Utilitarian point of view” (Sidgwick 1874). This has since been echoed by the likes of Jonathan Glover (1977) and Parfit (1984), among others.

In assessing human extinction, then, such nuances are incredibly important. Consider one of the questions asked by the aforementioned survey by Coleman et al. They write that “we asked participants in both the U.S. and China whether human extinction would be good, bad, or neither” (Coleman et al., unpublished). For present purposes, let’s simplify this to the question of “whether human extinction would be bad.” As established in previous sections, this question is ambiguous in multiple ways: does “human” mean “*Homo sapiens*” or “*Homo sapiens* and whatever descendants we might have with the right sort of moral status”? Furthermore, what kind of “extinction” are we talking about? If one adopts the Narrow Definition, the term could refer to terminal, final, or phyletic extinction; if one adopts the Broad Definition, the term could refer to terminal or normative extinction. Without specifying how these terms are defined, the question is hopelessly unclear. So, let’s say—again, for present purposes—that the question concerns final extinction, given the Narrow Definition. Yet the question remains deeply problematic, since, *depending* on one’s preferred position in Existential Ethics, it very much matters *how* Going Extinct unfolds. When most people are asked to imagine our extinction, they will imagine this happening as a result of a catastrophe. A catastrophic etiology is *part* of what I will call the “prototypical conception” of human extinction, where the prototypical conception—the conception that comes to mind for most people when they think of “human extinction”—corresponds to the “fi-

nal extinction of *Homo sapiens* caused by a catastrophe,” such as an asteroid impact or a Terminator scenario.⁹ This etiological detail matters because most people will say that human extinction is bad *at least* because Going Extinct would, if caused by a catastrophe, cause lots of physical and/or psychological suffering. But most people, according to one study, also share person-affecting intuitions that lead them to argue that the bigger difference in Parfit’s thought experiment is between (1) and (2), not (2) and (3). Hence, on the assumption that human extinction is caused by a catastrophe, answers to the question above might give the impression that there is wide agreement about the badness/wrongness of human extinction, when in fact there are considerable disagreements just below the surface. If the question explicitly specified that the etiology of our extinction were that everyone around the world agreed to voluntarily stop having children, many people might give a different answer, thus revealing a deeper divergence in ethical and evaluative opinions about the non-existence of our species.

To make this a bit more concrete, if someone were to ask me whether the final extinction of *Homo sapiens*, caused by a catastrophe, would be bad, I would answer: “Yes, absolutely.” This is the same answer that, say, longtermists and totalist utilitarians would give, which suggests that I am aligned with the views of these two groups. However, I am an equivalence theorist, and hence do not care *at all* about *extinction itself*. In my view, the badness of the situation in figure 5 plateaus once the casualties reach 100%. I care about people not suffering, and hence believe that we should take measures to prevent our extinction *if caused* by a catastrophe; if the cause is voluntary, uncoerced, peaceful, and so on, then I do not have a problem with humanity going extinct. These crucial differences are only clear if one understands the polysemy of “human extinction” and the distinction between Going Extinct and Being Extinct. This is why the claims of this paper are of vital importance for future research on this topic, whether scientific or philosophical.

Yet we have neglected thus far to fully register a third major position within Existential Ethics, which I am also sympathetic with: *pro-extinctionism*. The most common interpretation of this position makes at least the follow specific claim: Being Extinct would be in some sense *better than* Being Extant, i.e., continuing to exist (see Redacted for a detailed discussion). This does not mean that Being Extinct would be *good*. For example, Simon Knutsson argues that “an empty world is the best possible world,” but he adds that “I would not say that an empty world would be good” (Knutsson 2023). Alternatively, Benatar seems to believe that Being Extinct would be positively good, because it would mean the absence of both pleasure and pain, where the absence of pleasure is not bad and the absence of pain is good. This contrasts with the situation of existence, which involves the presence of both pleasure and pain, where the presence of pleasure is good and the presence of pain is bad—i.e., a good/bad situation. A not-bad/good situation (non-existence) is not only better than a good/bad one (existence), but *positively* good (Benatar 2006).

⁹ To be clear, the prototypical conception is an empirical hypothesis. I have considerable anecdotal evidence that most people understand “human extinction” as the “final extinction of *Homo sapiens* caused by a catastrophe,” but this hypothesis lacks good scientific evidence. I would be eager for psychologist to prove me right or wrong.

A key point to make with respect to pro-extinctionism is that it is, in nearly all cases, a claim about Being Extinct versus Being Extant, not about Going Extinct.¹⁰ Many pro-extinctionists agree that if Going Extinct involves lots of suffering and/or cuts lives short, it would be very bad or wrong. This is why the majority of pro-extinctionists from the German pessimists of the latter 19th century up to the present have strongly *opposed* *omnicide*, or “the murder of everyone.” Benatar, for example, distinguishes between a “dying-extinction” and a “killing-extinction,” where the former is, roughly speaking, voluntary while the latter is not. (Omnicide would thus be one type of killing-extinction, though there could be non-anthropogenic causes as well.) He then argues that the only morally acceptable way of bringing about our extinction—and here, again, he’s thinking about final extinction on the Narrow Definition, or terminal extinction on the Broad Definition—would be voluntarily, by choosing not to have children.

To be clear about this point, pro-extinctionists confront a practical-ethical problem that is unique to their position: if Being Extinct is better than Being Extant (evaluative claim), and if Being Extinct is something that we ought to strive for (deontic claim), how do we get from here to there? How should we bring about our extinction? There are three main possibilities on the menu of options: *omnicide*, whereby one or more people kill everyone (or nearly everyone); what I will call *pro-mortalism*, whereby everyone (or nearly everyone) kills themselves; and *antinatalism*, whereby everyone (or nearly everyone) voluntarily chooses to stop procreating (and opts not to indefinitely extend their lifespan, if that option is available).¹¹ Most pro-extinctionists since the German pessimists, including pro-extinctionists motivated by ecological considerations, have identified antinatalism as the only permissible path to our extinction.¹² At the same time, most pro-extinctionists have agreed that the antinatalist path is extremely improbable, as there is almost zero chance that a sufficient number of people around the world will voluntarily decide not to have children, at least within the foreseeable future; pronatalist tendencies are simply too dominant. Hence, if humanity does go extinct, it will almost certainly be the result of a catastrophe, which many pro-extinctionists believe would be very bad or wrong.

¹⁰ Failing to appreciate this point can lead to misleading claims, as when the journalist Dylan Matthews writes that “unless you are a member of the Voluntary Human Extinction movement, you’ll probably agree that human extinction is indeed bad” (Matthews 2022). The Voluntary Human Extinction Movement (VHEMT), though, strongly opposes any form of human extinction that would be involuntary, especially if it were to entail physical or psychological suffering. Members of the community would contend that if an asteroid were barreling toward Earth, and if there were something we could do to redirect it away from Earth, we should redirect it—even though the outcome of Being Extinct would be better than Being Extant.

¹¹ I say “nearly everyone” because it is not necessary for, say, an omnicidal actor to kill everyone on Earth for humanity to disappear entirely. They would, however, need to kill enough people for the global population to dip below the “minimum viable population,” at which point *functional extinction* would be inevitable, where “functional extinction” refers to a scenario in which members of the species still exist, but the species’ extinction is nonetheless guaranteed. Note also that one could advocate for a combination of pro-mortalism and antinatalism, whereby some people kill themselves while others decide not to procreate. Omnicide, in contrast, is all or nothing, and hence cannot be combined with these other options.

¹² There are exceptions, though. Philipp Mainländer also suggested a pro-mortalist route, and indeed he committed suicide days after volume I of his *magnum opus* was published. Eduard von Hartmann opposed antinatalism, instead arguing (roughly speaking) that as civilization develops, a means for eliminating all life in the universe, including all human life, will gradually come into view. Even on this view, though, Hartmann seems to have believed that total annihilation in the future ought to be voluntary: the development of our consciousness over time, he claimed, will lead a growing number of people to realize that existence is very bad, and hence that nothing should exist. Because he was an idealist, he believed that if all subjects are eliminated from the universe, the universe itself will cease to be (see Beiser 2016, ch. 7).

Returning once more to figure 4, pro-extinctionists would say that in a forced-choice situation, the scenario of World B would be preferable, independent of the etiology. That is to say, 10 billion people in each world deciding not to have children would be ideal, but if the event in each of these worlds is a catastrophe that kills 10 billion people, at least there would be some silver lining to the second scenario: humanity would no longer exist, thus erasing all the future suffering that could have otherwise existed (in the case of philosophical pessimism) and preventing the further destruction of the environment by *Homo sapiens* or our descendants (in the case of radical antihumanist environmentalism). Hence, further-loss theorists see the scenario of World B as (much) worse than that of World A; equivalence theorists claim that there is no difference between the two scenarios; while pro-extinctionists will say that the scenario of World B is better, though most would strongly prefer that humanity dies out through some voluntary, peaceful means rather than because of a catastrophe.

6. Conclusion

This paper delineates several distinctions that, I argue, are crucial for making sense of Existential Ethics. As of now, the literature is a mess of conceptual, ontological, and terminological confusion, with many contributors failing to make clear how they are defining “human” and “extinction,” while also being unclear about the importance of Going Extinct and Being Extinct for assessments of the ethical and evaluative implications of our disappearance. While most people will, I believe, intuitively understand “human” as denoting “*Homo sapiens*,” many futurists use “human” and “humanity” to pick out a much larger class of beings. Indeed, the Broad Definition tends to fit more naturally with most further-loss and pro-extinctionist views, while equivalence views may be neutral between the Narrow and Broad definitions, since all that matters to such views is whether there is anything bad or wrong with the processes or events leading up to our extinction, however defined. Furthermore, different ethical and evaluative positions will identify different types of extinction as very bad or wrong, while simultaneously identifying other types as neutral or even good. Longtermists, for example, may see the phyletic extinction of *Homo sapiens* as a positive development, if the posthuman beings that we evolve into—through natural processes or cyborgization—are better able to fulfill “our longterm potential” than we are. Most further-loss views identify terminal and normative extinction, on the Broad Definition, as the only types of extinction that must be avoided; they would, therefore, also claim that the terminal extinction of our particular species, *Homo sapiens* (Narrow Definition), would not *in itself* be bad or wrong. Most pro-extinctionist views will claim that we should aim for Being Extinct in the final (Narrow Definition) or terminal (Broad Definition) senses, and most equivalence views will be indifferent about the type of extinction that we might undergo: whichever type of extinction occurs, if the process or event of Going Extinct is bad or wrong, then that type of extinction is bad or wrong; if Going Extinct is not bad or wrong, then there is nothing bad or wrong with that type of extinction.

My own view is that scholars should adopt, as standard practice, the Narrow Definition over the Broad Definition, as this would enable more nuance in discussions about human extinction, since the Narrow Definition foregrounds three distinct types of extinction scenarios, two of which the Broad Definition conflates (namely, final and terminal extinction). The Narrow Defini-

tion also makes clear that many transhumanist and longtermists—among the loudest voices talking about the importance of avoiding “human extinction” these days—are okay with, if not in favor of, the disappearance of *Homo sapiens* in the long term, or even the near term (later this century). Indeed, assuming the Narrow Definition, one should classify these individuals as *pro-extinctionists*, in the particular sense that they do not oppose our species disappearing entirely and forever so long as we are replaced by a “superior” new species of posthumans (for useful discussion, see Kirsch 2023 and Redacted forthcoming). What matters to such people is that *Homo sapiens* survives, for instrumental reasons, at least long enough to create our successors. As for normative extinction, this should be explicitly linked to the idea of posthuman beings who may or may not exist in the more distant future.

The present paper does not offer an exhaustive treatment of this subject. However, I do hope this provides some much-needed conceptual clarity about this issue, which is—sadly—of increasing relevance given the worsening climate crisis, ongoing wars involving nuclear-armed countries, and the recent development of powerful AI systems.

References

Beckstead, Nick. 2013. “On the Overwhelming Importance of Shaping the Far Future.” PhD Dissertation. Rutgers the State University of New Jersey-New Brunswick. <https://rucore.libraries.rutgers.edu/rutgers-lib/40469/PDF/1/play/>.

Benatar, David. *Better Never to Have Been: The Harm of Coming into Existence*. Oxford: Oxford University Press.

Bostrom, Nick. 2005. “A Philosophical Quest for Our Biggest Problems.” TED. www.Ted.Com/Talks/Nick_bostrom_on_our_biggest_problems.html.

Bostrom, Nick. 2008. “Why I Want To Be a Posthuman When I Grow Up.” In Bert Gordon and Ruch Chadwick (editors), *Medical Enhancements and Posthumanity*. New York, NY: Springer.

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

Bostrom, Nick, and Toby Ord. 2006. “The Reversal Test: Eliminating Status Quo Bias in Applied Ethics.” *Ethics*. 116(4): 656-679.

Coleman, Matthew, Lucius Caviola, Joshua Lewis, and Geoffrey Goodwin. Unpublished. “How Important Is the End of Humanity? Lay People Prioritize Extinction Prevention but not Above All Other Societal Issues.” <https://osf.io/preprints/psyarxiv/qn7k5/download>.

Delord, Julien. “The Nature of Extinction.” *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*. 38(3): 656-67.

Frick, Johann. 2017. “On the Survival of Humanity.” *Canadian Journal of Philosophy* 47, no. 2–3 (2017): 344-67.

Gibbons, Ann. 1993. “Pleistocene Population Explosions: A Controversial Method of Reconstructing PreHistorical Populations Indicates that Separate Modern Human Groups—and not a Single Group from Africa—Suddenly Expanded about 50,000 Years Ago.” *Science*. 262(5130): 27-28.

Goertzel, Ben. 2010. *A Cosmist Manifesto*. Humanity+ Press. https://goertzel.org/CosmistManifesto_July2010.pdf.

Glover, Jonathan. 1979/1990. *Causing Death and Saving Lives: The Moral Problems of Abortion, Infanticide, Suicide, Euthanasia, Capital Punishment, War and Other Life-or-Death Choices*. New York, NY: Pelican Books.

Greaves, Hilary, and William MacAskill. 2021. “The Case for Strong Longtermism.” *Global Priorities Institute*. <https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf>.

Jonas, Hans. 1979. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Mersion: Emergent Village Resources for Communities of Faith Series. Chicago, IL: University of Chicago Press.

Kirsch, Adam. 2023. *The Revolt Against Humanity: Imagining a Future Without Us*. New York, NY: Columbia Global Reports.

Leiserowitz, Anthony, Edward Maibach, Connie Roser-Renouf, Seth Rosenthal, and Matthew Cutler. “Climate Change in the American Mind.” *Yale Program on Climate Change Communication and George Mason University Center for Climate Change Communication*, 2017. <https://climatecommunication.yale.edu/wp-content/uploads/2017/07/Climate-Change-American-Mind-May-2017.pdf>.

MacAskill, William. 2022. *What We Owe the Future*. New York, NY: Basic Books.

Matheny, Jason. 2007. “Reducing the Risk of Human Extinction.” *Risk Analysis: An International Journal*. 27(5): 1335-44.

Mecklin, John. 2024. “2024 Doomsday Clock Statement.” *Bulletin of the Atomic Scientists*. <https://thebulletin.org/doomsday-clock/current-time/>.

Merriam-Webster. 2021. “Extinct.” https://www.merriam-webster.com/dictionary/extinct?utm_campaign=sd&utm_medium=serp&utm_source=jsonld.

MUP. Most Expect ChatGPT Will Be Used for Cheating. Monmouth University Poll. February 15, 2023. https://www.monmouth.edu/polling-institute/reports/monmouthpoll_us_021523/.

Narveson, Jan. 1967. "Utilitarianism and New Generations." *Mind*. 76(301): 62-72.

Ord, Toby. 2020. *The Precipice: Existential Risk and the Future of Humanity*. New York, NY: Hachette Books.

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

Rees, Martin. 2003. *Our Final Hour*. New York, NY: Basic Books.

Sandberg, Anders, and Nick Bostrom. 2008. "Global Catastrophic Risks Survey." Future of Humanity Institute, Technical Report #2008-1. <https://www.fhi.ox.ac.uk/reports/2008-1.pdf>.

Scheffler, Samuel. 2007. "Immigration and the Significance of Culture." *Philosophy & Public Affairs*. 35(2): 93-125.

Scheffler, Samuel. 2018. *Why Worry about Future Generations?* Uehiro Series in Practical Ethics. Oxford: Oxford University Press.

Schubert, Stefan, Lucius Cavil, and Nadira Faber. 2019. "The Psychology of Existential Risk: Moral Judgments about Human Extinction." *Scientific Reports*. 9, 15100.

Shelley, Mary. 1826/2008. *The Last Man*. Oxford World's Classics. Oxford: Oxford University Press.

Sidgwick, Henry. 1874. *The Methods of Ethics*. Donald F. Koch American Philosophy Collection. Macmillan. <https://books.google.de/books?id=KVAtAAAYAAJ>.

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

Thomas, Alexander. 2022. "The Politics and Ethics of Transhumanism: Exploring Implications for the Future in Advanced Capitalism." PhD Dissertation. University of East London. https://repository.uel.ac.uk/download/8a1d0d464e1f89ded4ace911c3470e192-da96f564aaae7c68e1c787fc2ca6a23/931830/2022_PhD_Thomas.pdf.

Thomas, Alexander. Forthcoming. TBD. (Academic press.)

Tonn, Bruce. 2009a. "Obligations to Future Generations and Acceptable Risks of Human Extinction." *Futures* 41(7): 427-35.

Tonn, Bruce. 2009b. "Beliefs about Human Extinction." *Futures*. 41(10): 766-773.

Vetter, Hermann. 1968. "Discussion." In Paul Weingartner and Gerhard Zecha (editors), *Induction, Physics, and Ethics*. Dordrecht, Netherlands: D. Reidel Publishing Company.