How Tragic Would Human Extinction Be?
Convergent Arguments for Making the Survival of Our Lineage a Global Priority

Abstract: This paper synthesizes a wide range of distinct moral and axiological arguments that all converge upon the evaluative conclusion that human extinction, if it were to occur, would constitute an immense tragedy. In doing so, it affirms that one need not champion any particular moral or axiological position to agree that we should make the avoidance of extinction a global priority. We examine five features of human extinction that could make our disappearance tragic, and then explore how four classes of moral theories, which together cover the vast majority of positions within contemporary ethics, would evaluate these features. Finally, we consider the possibility that humanity does not yet know what the correct conception of morality is, and thus does not currently understand how tragic our extinction might be. This makes the prospect of our imminent disappearance even worse, because it suggests that we might never know just how valuable we were or how important our survival is.

Section 1. Introduction

> *Without the possibility of a future, there is nothing left but despair. Thus, if we give up on the future, we give up on ourselves.—Wendell Bell*[1]

> *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. In Nietzsche's words, there has never been such a new dawn and clear horizon, and such an open sea.—Derek Parfit*[2]

Would it be wrong if humanity were to go extinct like most species that have so far existed? If so, how wrong, and for what reasons? The present paper aims to synthesize a wide range of arguments for why human extinction would be *very bad*, if not *one of the worst things that could possibly happen*. We will call this "Conclusion C." Although there are some moral and axiological positions that see the extinction of humanity as desirable, such as Benatarian anti-natalism, our aim will be to show that there are multiple independent lines of reasoning that all converge upon Conclusion C, and thus imply that humanity ought to make the avoidance of extinction a top global priority this century and beyond. Here it is useful to distinguish between (E) the event or process of *going extinct*, and (S) the state or condition of *being extinct*. At the very least, all of the most prominent moral theories identify (E), if caused or allowed by moral agents, as wrong; other theories, however, affirm the moral badness of both (E) and (S), even going so far, in certain cases, to assert that *most* of what is bad about human extinction is the subsequent loss of value that could have been realized if only we had survived.

This topic is not one of mere philosophical curiosity. As one of us has elsewhere shown, the very idea of *human extinction* is a quite recent addition to our shared conceptual repertoire; e.g., not even Charles Darwin entertained the idea in his work on human evolution.[3] Consequently, very little scholarly attention has focused on the ethical implications of either (E) or (S). Yet the most informed probability estimates of human extinction occurring this century suggest that this could be the most dangerous moment of our species' 200,000-year history. For example, Nick Bostrom (2005) estimates that the likelihood of extinction before 2100 is not less than 20 percent; an informal survey of experts at the Future of Humanity Institute (FHI) yields a median probability of 19 percent (Sandberg and Bostrom 2008); and Toby Ord conjectures a 1-in-6 chance of extinction before 2100 (see Wiblin 2017; author). For ease of discussion, we can call this the "unique hazards hypothesis," defined as follows:

---

[1] See Bell 1993. This quote also appears in author.
[2] This passage appeared on the final page of Parfit 2017, published shortly after the author's death. A fuller version of the Nietzsche quote was included at the beginning of his first book, Reasons and Persons. "At last the horizon appears free to us again, even granted that it is not bright; at last our ships may venture out again, venture out to face any danger; all the daring of the lover of knowledge is permitted again; the sea, our sea, lies open again; perhaps there has never yet been such an 'open sea.'" (See Nietzsche 1977, 488.)
[3] To be clear, Darwin did anticipate human annihilation as a result of the second law of thermodynamics, but he never considered it in the context of evolutionary biology.

Unique hazards hypothesis: Humanity finds itself in a period of historically unprecedented dangers to our survival.

If this hypothesis is even remotely accurate, then the question of whether and to what extent Conclusion C is well-supported by cogent philosophical argumentation is not only of paramount importance but extremely urgent as well.[4] Thus, the present paper aims to fill-in a significant lacuna in the contemporary literature on existential risks, offering a robust point-of-departure for future discussions of whether and to what extent human extinction would be tragic.

This paper will proceed as follows: Section 2 examines five potential reasons why human extinction might be tragic. Section 3 explores how the four main classes of moral theories would evaluate these reasons. Section 4 addresses the issue of normative uncertainty in the context of human extinction. It should become clear by the end of this fairly exhaustive tour through a philosophical labyrinth of argumentation that one need not espouse any particular moral or axiological perspective to see human extinction as a tragedy of immense proportions. Indeed, some lines of reasoning that converge upon Conclusion C are not only independent but mutually exclusive, meaning that people with incompatible moral or axiological commitments can still agree that ensuring our survival into the far future should constitute a global priority for human civilization.

Section 2. Five Potential Tragedies of Human Extinction

There are at least five aspects of human extinction that moral and axiological theories could appeal to in arguing that such an event would constitute a tragedy of the highest order. These are:

2.1 *Human extinction will very likely harm those alive at the time*. Although no scientific surveys of normative beliefs about human extinction have been conducted, there are reasons for suspecting that most people would pre-theoretically see (E) as tragic for at least one morally relevant reason, namely, that it would entail the death of ~7.6 billion people (as of this writing); yet many would also see this as only slightly worse than a catastrophe that kills a much smaller number of people. Even more, most probably wouldn't view (S) as being bad at all.[5] Taking these in reverse order, many philosophers and non-philosophers alike hold "person-affecting" intuitions according to which we should care most, or only, about the effects of our actions on currently existing people and consequently little, or not at all, about those who will only exist in the future. Thus, since no one will be around to bemoan the non-existence of humanity, who cares? On the other hand, while many agree that any bodily or psychological harm that going extinct causes would be bad, the badness of such harm does not scale multiplicatively as the number of deaths increases due to cognitive biases like "scope neglect" and "psychophysical numbing" (see Slovic 2017). The first bias refers to our inappropriate emotional responses to large numbers and the second denotes the rapid decline in compassion for other humans as the absolute number of casualties rises above *one*. Thus, most people interpret the difference between 0 and 1 deaths as *greater* than the difference between 2,154,489,204 and 2,154,489,205 deaths; most people would rather spend some quantity of resources to prevent a single death rather than to prevent 2,154,489,204 deaths from becoming 2,154,489,205. Yet there is no reason why this should be carried over into our moral evaluations as such, and once we evaluate human extinction according to the *actual* number of people who will die it seems that, even if we do hold person affecting views, an extinction event would still be one of the worst things that could ever happen.

Furthermore, even if people preserve such biases, there remains the fact that an event of this sort could, depending on the timing, cause the death of oneself or living creatures that one cares about, such as parents, siblings, friends, pets, or celebrities. Consider that if the risk of human extinction were only 0.1 percent per year—a relatively conservative estimate—and if the only life one cared about was one's own, then one should *still* be willing to take precautions against such extinction that are at least as costly as

---

[4] Note here that human extinction is probably the *least probable* existential risk "failure mode," given the "last few people problem" that one of us discusses elsewhere (see author).

[5] Cf. Scheffler 2018.

those one takes against dying in a road traffic accident, since the chance of death from both causes would be comparable (Global Challenges Foundation 2017; author)[6] But death itself isn't the only personal harm that the scenario above could entail. People often seek to acquire a form of "vicarious immortality" by contributing in some way to culture, sports, academia, and so on; the thought of such accomplishments, or "traces," persisting through time give many people a deep sense of meaning in life. Or, as Wilhelm Ostwald's put it in 1906, "every man leaves after his death certain things in the world [that are] changed by his influence," adding that "there is a very general desire in mankind to leave such impressions."[7] Along these lines, Ernest Partridge (1981) argues that human beings manifest a

> desire to extend the term of one's influence and significance well beyond the term of one's lifetime—a desire evident in arrangements for posthumous publications, in bequests and wills, in perpetual trusts (such as the Nobel Prize), and so forth. In such acts and provisions, we find clear manifestations of a will to transcend the limits of personal mortality by extending one's self and influence into things, associations, and ideals that endure.[8]

This gestures at yet another idea that Janna Thompson (2009) calls "lifetime-transcending concerns," or "interests concerning states of affairs that will, or could, occur in the future beyond one's lifetime." It follows that merely *believing* that humanity will go extinct in the foreseeable future could lead to a sense of despondency by undercutting the promise of vicarious immortality. In the words of Allen Tough (1991), if humanity fails to survive, "then most other values and goals will lose their point. No other goals are more important than humanity's survival at a satisfactory level." Elsewhere he writes that "if humanity goes out of existence or back to the caves, then our personal efforts and achievements also disappear. Career success, national prestige, battlefield victories, business as usual, books, paintings, and children will not provide any of us with long-lasting benefits if human civilization destroys itself."[9] Thus, not only would human extinction entail actual mortality, but it would also eliminate vicarious immortality.

2.2 *Human extinction will cause a great loss in the quantity of human welfare and other values*. Many of the scholars and philanthropists who are most committed to averting human extinction are motivated not only by the harm of (E), but also by considerations of the lost value inherent in (S). There are several possible routes to this conclusion depending on one's interpretation of the *astronomical value thesis*:

> *Astronomical value thesis*: The potential value of the future could be astronomically huge, if only we play our cards right.

The question immediately arises as to what "value" means, and one answer is: "Whatever you would like it to mean." The point is that if humanity survives and spreads through the universe, there could be far more of *whatever one values* in the future than in the contemporary world. In fact, philosophers have argued that, insofar as one genuinely values a property P, one should strive to ensure its continued existence into the future, if not work to multiply the number of instances of the universal (see Scheffler 2007, 2018).[10] It follows that one should see human extinction as bad not just because of the possible bodily and psychological harm that going extinct could cause, but because of the loss of potentially astronomical

---

[6] Indeed, as James Wilhelm (2012) points out, there could be many tokens of the "vulgar egoist who claims that that which occurs after his life has ended is unimportant" type.

[7] Quoted in Trisel 2004.

[8] Quoted in *ibid*.

[9] The same point was also made by David Collings in 2014, who, in the context of discussing climate change, asks: "What actions do we take that do *not* imply a relation to the future? What would be the point of our lives if the future were to disappear? … *The future is never just for the people of the future; without that future, what we do now loses its force*. Without a future, there is no present and not much of a past. Climate change isn't just about our obligation to others. It's about our *own* lives, too. … If we give up on climate change, we cancel the present. We choose to make all of our ordinary activities meaningless, as if we want to become shadows of ourselves, or as if we want to float forever in a world without foundations."

[10] See Lenman 2002 for elaboration and criticism of the latter point.

amounts of value that we could actualize.[11] In this way, given what it *means* to value something and that people do value thing, the descriptive proposition of the astronomical value thesis yields the normative proposition of Conclusion C.

A popular moral interpretation of the axiological component of the astronomical value thesis comes from what is called the "Total View."[12] This identifies value as a property of individual lives, where "value" can be defined in very many ways including hedonistic, desire-satisfaction, or objective list-theoretic terms. The central idea is that morally good acts are those that increase the total amount of value across all lives, present and future. A weaker interpretation of "astronomical value" arises from what is called the "Simple View." On this view, value is still a property of individual lives, but one is not necessarily committed to maximizing it. Rather, all this view implies about morally good acts is that some lives are "good" and that the addition of any of these lives would always be in itself good and make the world better in at least one way.[13] It follows that, on either of these views, since human extinction would permanently preclude the realization of potentially vast amounts of future value, "to end the human race would be about the worst thing it would be possible to do," a sentiment that goes back at least to Henry Sidgwick (1907). From this perspective, the difference between 99 and 100 percent of humanity dying out is *far greater* than the difference between 1 and 99 percent perishing (Parfit 1984). The reason is that, as H.G. Wells—the founder of future studies[14]—put it in 1902, is that

> all the past is but the beginning of a beginning, and that all that is and has been is but the twilight of the dawn. It is possible to believe that all that the human mind has ever accomplished is but the dream before the awakening. We cannot see, there is no need for us to see, what this world will be like when the day has fully come. We are creatures of the twilight.

Indeed, according to Carl Sagan, if humans remain on Earth with an average lifespan of 100 years, there could come to exist 500 trillion humans in the future. But it appears likely that, if we survive the next few centuries, humanity will spread into the cosmos, which could vastly increase the total number of people—and thus the total amount of well-being in the universe. For example, Bostrom (2003b) calculates that if a single star can sustain ~10 billion people, then the Virgo Supercluster could house ~100 sextillion future humans per century; yet there are about 10 million superclusters in the observable universe and 1 billion trillion stars in total. (We will leave it to readers to do the math!) Even more, if mind-uploading is possible, then based on calculations of the computational capacity of planets that are converted into giant supercomputers, there could exist ~100 decillion (or 10^{38}) people *per century* in the Virgo Supercluster alone, although Milan Ćirković (2002) puts the number even higher at 10 quattuordecillion (or ~10^{46}). If such people have worthwhile lives, then the potential well-being that our descendants could realize in the future could be truly astronomical.[15] Notice here that even if one highly discounts future lives, the sheer number that could come into existence still suggests that existential risk reduction ought to be highly prioritized.

---

[11] In other words, even if human extinction were to be painless and instantaneous, it would still constitute an immense tragedy because of *what could have been*.

[12] For a new and interesting alternative take, see Kaczmarek 2018.

[13] These two views are discussed in Parfit 2016.

[14] According to W. Warren Wagar 1983.

[15] It is worth noting here that the "big freeze" of some models of physical eschatology might not necessitate the annihilation of humanity; or, as Michio Kaku (2005) puts it, "the ultimate death of our universe may not necessarily mean the death of intelligence." The reason is that—and to be clear, this is highly speculative—a type III (or higher) civilization on the Kardashev scale could "use 'exotic energy' and black holes to find a passageway to another universe." If this were to fail, "there is another final option: To reduce the total information content of an advanced, intelligent civilization to the molecular level and inject this through the gateway, where it will then self-assemble on the other side. In this way, an entire civilization may inject its seed through a dimensional gateway and reestablish itself, in its full glory. Hyperspace, instead of being a plaything for theoretical physicists, could potentially become the ultimate salvation for intelligent life in a dying universe." Thus, there is at least a prima facie reason for thinking that the amount of value that we could realize in the future might not be merely astronomical but multi-universal—or perhaps even genuinely infinite—if we become parallel-universe nomads.

2.3 *Human extinction will cause a great loss in the quality of human welfare and other values.* The astronomical waste thesis, though, does not need to be understood solely in terms of the total number of people with worthwhile lives that could come to exist in the future. Indeed, it should also take account of the potential for people in the future to acquire lives that are not merely worthwhile but *extraordinarily*, perhaps *unimaginably* (from our current vantage point), good. There are two trends worth mentioning here, the first being *societal*: As Steven Pinker (2011) outlines in great detail, humanity has made significant moral progress throughout history, and especially since the end of World War II. One happy symptom of this is the steady, if uneven, decline of violence. According to Pinker, humanity is not only in the midst of the "Long Peace" but also the "New Peace," during which organized conflicts of all kinds—civil wars, genocides, repression by autocratic governments, and terrorist attacks—[have] declined throughout the world." Although Pinker does not extrapolate these trends into the future, there are some reasons for expecting them to continue. For example, Pinker identifies the primary driver of moral progress in recent decades as the Flynn effect, the long term increase in human intelligence observed during the twentieth century. While this appears to have slowed, stopped, or reversed in certain regions of the world (see, e.g., Bratsberg and Rogeberg 2018), it is very likely that future innovation will yield safe and effective cognitive and moral enhancements. If the former augments our capacity for "abstraction from the concrete particulars of immediate experience," which is "the cognitive skill that is most enhanced in the Flynn effect," then we might expect a further expansion of our circles of moral concern (Pinker 2011). There could be similar gains from moral bioenhancement, which Ingmar Persson and Julian Savulescu (2012) describe as any biomedical intervention that augments our moral dispositions of altruism and a sense of justice (or fairness). The first consists of empathy—the cognitive property that Pinker focuses on—and sympathetic concern—the motivational element of moral action. Although Persson and Savulescu's proposal has proven to be controversial for both philosophical and practical reasons (see author), it is not unthinkable that future breakthroughs yield highly effective *mostropics*, i.e., morality-boosting drugs, and that these drugs become as commonly ingested as fluoride is via the public drinking water.

This leads to the second trend, which is *transhumanist*: Human enhancement technologies could enable a phase transition from humanity to posthumanity, where posthumans are beings with significantly augmented capacities in the broad domains of cognition, emotion, and healthspan (see Bostrom 2008). For example, nootropics, transcranial magnetic stimulation, brain-computer interfaces (BCIs), genetic modifications, iterated embryo selection, mind-uploading, *and so on*, could potentially increase our intellectual abilities, while advanced biotechnology and molecular nanotechnology could stop or even reverse aging, thus enabling people to live indefinitely long lives.[16] The result could be a population of beings who experience degrees of well-being that far exceed the intensity and amount that any current human could possibly attain. Even more, enhancement technologies could expand our "cognitive space" such that our posthuman progeny have mental access to concepts that are in principle beyond our ken (see author). It follows that, insofar as (say) some theory T requires a concept C to understand, and insofar as C falls outside of our current cognitive space but within the cognitive space of a species of posthuman, then that species could devise T—a theory about which we might be second-order ignorant, meaning that we can't even know that we can't know T. Thus, insofar as one values knowledge, there could be any number of marvelous new ideas in the future that forever linger beyond *our* epistemic reach.

This being said, people have been speculating about utopian futures in which people live qualitatively (much) better lives since at least the nineteenth century, and it has become natural to view these with a strong degree of skepticism. Yet 200 years ago, very few individuals were permitted the freedom of thought and self-expression that many now take for granted; violence, sickness, and disability where far more common than they are today; people's understanding of the world around them was quite impoverished and their access to science, art, and technology almost unimaginably lower than such access is today. Even if twenty-first century utopian visions remain unrealized, the prospects for a qualitative im-

---

[16] Below we will discuss the possibility of achieving a kind of "vicarious immortality" through one's life works. However, the following Woody Allen quote immediately comes to mind: "I don't want to achieve immortality through my work; I want to achieve immortality through not dying. I don't want to live on in the hearts of my countrymen; I want to live on in my apartment." Life-extension technologies could someday enable people to the sort of *actual* immortality for which Allen longs.

provement in future people's lives, we would argue, ought not be underestimated. Thus, there both quantitative and qualitative aspects to, at the very least, a Totalist interpretation of the astronomical value thesis.

2.4 *Human extinction will remove our rational/moral agency from the earth/universe*. The physicist Enrico Fermi once observed that the sheer size of the universe should entail that interstellar civilizations are relatively common, and we should be able to detect them with our current technology, however so far we have found no real evidence of their existence. This is known as the "Fermi Paradox." Yet there is a growing body of research that resolves (or dissolves) the paradox by arguing that intelligent life could be rare in the universe; at the extreme, humanity could be something of a cosmic *hapax os*, i.e., a "thing existing only once."[17] Peter Ward and Donald Brownlee (2000), for example, defend a version of the "rare Earth hypothesis," and a more recent analysis calculates that there is a 38 to 85 percent chance that humanity is alone in the observable universe and a 53 to 99.6 percent chance that humanity is alone in our galaxy (Sandberg et al. 2018). It follows that, as Toby Ord declares, "it's very possible that we might be the most amazing and rare part of the whole universe, the only part of the universe capable of understanding the universe itself and appreciating its wonders" (quoted in Wiblin 2017).

For some, this observation provides additional support for the notion that humanities extinction would be a tragedy of cosmic proportions. As Derek Parfit (2016) argues,

> if we are the only rational beings in the Universe, as some recent evidence suggests, it matters *even more* whether we shall have descendants or successors during the billions of years in which that would be possible. Some of our successors might live lives and create worlds that, though failing to justify past suffering, would give us all, including some of those who have suffered, reasons to be glad that the Universe exists (emphasis added)[18]

Here one need not believe that each and every future human life is valuable to be intellectually (and emotionally) moved by this general point: It would *be a great shame* if humanity (a) were "one of a cosmic kind," so to speak, and (b) were to stumble into the eternal grave of extinction. Thus, the state or condition of being extinct would itself be bad because it would entail the absence of presence of something truly special—us.

According to Christine Korsgaard, one can distinguish between (a) valuing something for its own sake, or as an end in itself, in which case it has "final value," and (b) valuing it for the sake of something else, or as a means to an end, in which case it has only "instrumental value" (Korsgaard 1983, 170). Valuing the continued existence of humanity because of the quantitative or qualitative difference that human survival could make to "whatever one values," as discussed in the two proceeding subsections, can be said to concern only the instrumental value of the valued objects. However, some philosophers have argued that future generations of humanity also have final value, which would be lost if humanity where to disappear. This idea has been defended most recently by Johann Frick (2017). The argument, as Frick presents it, begins with the observation that people commonly attribute final value to a range of phenomena, such as languages, species, and cultures. This suggests "that humanity too, with its unique capacities for complex language use and rational thought, its sensitivity to moral reasons, its ability to produce and appreciate art, music, and scientific knowledge, its sense of history, and so on, should be deemed to possess final value." Assuming that this is the case, humanity should work to safeguard its survival since, quoting Samuel Scheffler (2007), "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?" Thus, if humanity is finally valuable, then the loss of a sufficiently large number of instantiations of the universal *humanity* "would be very bad, indeed one of the worst things that could possibly happen" (Frick 2017). Once more, this goes beyond claims about the badness of (E) and seems to identify (S) itself as highly undesirable, since a world A that has but then loses a finally valuable property P is less good than a world B that has and then retains P. Notice, though, that in this case the badness of (S) is far less counterfactual than in some of the cases above, as it reflects the inherent value of our nature rather than what we do with it.

---

[17] On the model of "*hapax legomenon*."

[18] See also Kahane 2014.

2.5 *Human extinction will end the narrative of humanity*. Finally let us consider aspects of human extinction that relate neither to the process of going extinct nor to what would be lost by the loss of humanities future, but to what extinction would mean for the history of humanity as a whole. These can usefully be broken down into three related groups.

2.5.1 *The argument from termination*. This begins with the observation that "it matters" how stories end; in particular, the anticipated narrative of civilization and humanity can affect the meaningfulness of the present moment. By analogy, consider someone reading a work of fiction: At any point in the book, one's emotional responses to the story will depend in part on how one expects the story to end. For example, when circumstances are favorable to the protagonist, the reader might feel a sense of hope, whereas when difficulties mount, the reader might feel a sense of dread and anxiety. The same idea applies to the grand narrative of *our species*, which is, from a naturalistic perspective, unfolding in a universe that lacks any intrinsic meaning or purpose. Thus, *believing* that human civilization will collapse, say, 100 years in the future and never recover could lead to a certain despondency in the here and now. This goes beyond "traces" and "vicarious immortality," arising instead from an abstract, impersonal desire for the human story to *end well*. Or, as Joshua Seachris (2011) puts it, "the way a narrative ends has significant proleptic [or retroactive] power to elicit a wide range of broadly normative human responses on, possibly, emotional, aesthetic, and moral levels towards the narrative as a whole *in virtue of it being the ending*."

Jonathan Bennett (1978) gestures at this idea when he asserts (discussed more below) that it would be "a pity, a shame, too bad, if the *story* were to be cut short in [our] world," as does Bostrom (2013) when he writes that "we ought not to delete or abandon the great *epic* of human civilisation that humankind has been working on for thousands of years, when it is clear that the *narrative* is far from having reached a natural terminus" (italics added). The critical point, though, is that we are not mere passive readers of a prewritten book—our biography—but active participants in a story that we ourselves are writing in realtime—a story whose ending ultimately depends upon our own individual and collective actions in the world.

2.5.2 *The argument from intergenerational justice*. Another reason to consider the arc of human history in its entirety, from humanity's cradle to our grave, is that it may be relevant to certain questions relating to justice. For instance, justice may oblige us to overcome the problem of past suffering on the grounds that a universe that contains a net balance of happiness over suffering is better than one that contains a net balance of suffering over happiness. It follows that, if one believes that history so far has contained more misery than joy—a claim that would be endorsed by many philosophers, including Arthur Schopenhauer, David Benatar, and perhaps even the present authors—then one might argue that it is very important that we survive, since survival is a necessary condition for working to rectify this wrong—for "giv[ing] us all … reasons to be glad that the universe exists" (Parfit 2011). A universe that includes our actual history *plus* astronomical amounts of net future value is surely better than a universe that includes our history but no future beyond, say, 5 years from now. Thus, we should avoid human extinction for the sake of making at least *some* recompense for past violence, wars, and genocides.

Even more, perhaps justice obliges us to overcome not just past human suffering, but past and ongoing suffering that humans have caused to non-human organisms. Consider that we are the number one driver of global biodiversity loss, which has resulted in the sixth mass extinction in life's 3.8-billion-year history. As Jennifer Jacquet (2017) observes, humanity has probably changed the biosphere more than any other organism since cyanobacteria brought about the Great Oxygenation Event ~2.45 billion years ago. This yields a view according to which, to quote R. Slaughter (1994), "the global commons has been compromised by human activity and restorative actions are necessary." Thus, it behooves humanity to continue existing in order to undo some or all of the damage that we have inflicted upon the natural world, an act that could be accomplished given (a) ongoing attitudinal shifts about sustainability, as well as (b) future technological possibilities, such as carbon removal and de-extinction techniques, that could restore climatic and ecological systems. An even grander point stems from the fact that the sun will eventually sterilize this planetary spaceship, thus destroying all life on Earth. It follows that, since humans are the only species with the capacity to migrate to and terraform exoplanets in the foreseeable future, and if one values the continued existence of Earth-originating life, then one should work to ensure that humanity

doesn't go extinct so that Earth-originating life doesn't go extinct either.[19] Along these lines exactly, humans are the only species with the capacity to intervene in the Darwinian "struggle for existence" and reduce the truly enormous amount of suffering among wild animals. In fact, this is the central aim of David Pearce's "abolitionist project," which imagines a world in which "our descendants will be animated by gradients of genetically preprogrammed well-being that are orders of magnitude richer than today's peak experiences" (Pearce 2007). Thus, even if one believes that humanity has so far been a great evil in the world—towards ourselves and other organisms—if one also believes in moral progress and cares about "cosmic justice," so to speak, then one could still come to the conclusion that we ought to prioritize our long-term survival.

2.5.3 *The argument from unfinished business*. Finally, human extinction may violate our responsibilities not only to the future and the timeless universe, but also to our past. Consider Edmund Burke's (1790) description of society as "a partnership not only between those who are living, but between those who are living, those who are dead, and those who are to be born." This suggests that if humanity were to go extinct, thus resulting in the termination of this partnership, something *extra* would be lost beyond the loss of humanity. It also implies that, insofar as past people pursued artistic, scientific, humanistic, athletic, religious, and so on, projects for the purpose of contributing to this partnership across time, then either causing or allowing human extinction to occur would be disrespectful to their posthumous wishes and interests. In fact, Thompson (2009) argues that desires that transcend our lifetimes could potentially "ground duties in respect to past people. If, for example, a person makes a request while alive about how her body would be disposed of, most people would probably agree that this gives her survivors a duty to respect her wishes (unless there is a good reason not to do so)." One could similarly argue, as Bennett (1978) does, that "one might assign (dis)utilities to past people," and this generates reasons to, e.g., "[object] morally to using the calculus for military purposes because Leibniz wanted all of his discoveries to contribute to universal peace." In other words, "if I use the calculus in building a bomb, I am bringing a disutility to Leibniz by bringing it about that he was to that extent a man whose hopes were not going to be realized."[20] The same logic could thus apply to the causing or allowing of human extinction, given that some past people have indeed wished for humanity to survive for as long as cosmologically possible.

The notion that civilization is a cooperative intergenerational partnership suggests, one could argue, a particular progressionist conception of history whereby humanity is "building something" over time that would be unfortunate to lose by being ended before the project is brought to completion. As Wendell Bell (1993) puts the point, "there is a *prima facie* obligation of present generations to ensure that important business is not left unfinished," where "important business" signifies "human accomplishments, especially exceptional ones in science, art, music, literature, and technology, and also human inventions and achievements of organizational arrangements, political, economic, social, and cultural institutions, and moral philosophy." Similarly, Robert Adams (1989) argues that

> the future of humanity [is] a vast project, or network of overlapping projects, that is generally shared by the human race. The aspiration for a better society—more just, more rewarding, and more peaceful—is a part of this project. So are the potentially endless quests for scientific knowledge and philosophical understanding, and the development of artistic and other cultural traditions. This includes the particular cultural traditions to which we belong, in all their accidental historic and ethnic diversity. It also includes our interest in the lives of our children and grandchildren, and the hope that they will be able, in turn to have the lives of their children and grandchildren as projects. To the extent that a policy or practice seems likely to be favorable or unfavorable to the carrying out of this complex of projects in the nearer or further future, we have reason to pursue or avoid it.

Bostrom (2013) makes a similar point in arguing that "we might also have custodial duties to preserve the inheritance of humanity passed on to us by our ancestors and convey it safely to our descendants. We do

---

[19] Thanks to Bruce Tonn, personal communication, for calling our attention to this.

[20] Bennett (1978) adds that "it seems that this would have been acceptable to Aristotle, who said that someone's 'happiness'—according to the standard translation—may be affected by what happens after his death."

not want to be the failing link in the chain of generations." Even more, Bruce Tonn contends that as humanity gets closer and closer to "completing its business," as it were, the stakes become exponentially higher, just as the tragedy of collapsing on the 25th mile of a ~26-mile marathon is greater than the tragedy of tripping on mile one amidst the pack. In Tonn's (2009) words, "humanity will suffer regret if it becomes extinct before its business in this Universe is finished and that this regret will eventually exponentially increase as the year one billion draws near if humanity's business is still unfinished."[21] There will, of course, not be anyone around to experience this regret if human extinction were to occur, but this does not entail that it wouldn't be substantively *regretful* from an impersonal perspective. It may be worth adding here that the idea of unfinished business partially motivates, insofar as it is motivated by any fundamental ideas, the "pro-humanity stand" of Bennett (1978), which we believe is worth mentioning. According to Bennett, this refers to

> a practical attitude … for which I have no basis in general principles. The continuation of *Homo sapiens*—if this can be managed at not too great a cost, especially to members of *Homo sapiens*—is something for which I have a strong, personal, unprincipled preference. I just think it would be a great shame—a pity, too bad—if this great biological and spiritual adventure didn't continue: It has a marvelous past, and I hate the thought of its not having an exciting future.

While not everyone will share Bennett's brute aversion to human extinction, we suspect that many people would concur with him, upon reflection, about the triumphs of human ingenuity and creativity. The irreversible foreclosure of an "exciting future" would be analogous to a young genius at the apotheosis of her career leaving a groundbreaking tome only half finished because she was killed by a stray bullet. In this case, her death would be tragic not only because it would entail death, but because it would leave some business unfinished, her virtues and capabilities being only partially realized.

Section 3. Evaluating The Tragedies of Human Extinction from Four Moral Perspectives

In the previous section, we delineated five features of human extinction that might contribute to its status as a moral or axiological tragedy. However, this doesn't provide a full account of why human extinction could be wrong. The present section thus aims to complete the analysis by exploring how different kinds of moral theories would evaluate the aforementioned features. In doing this, it also highlights the convergent nature of multiple arguments from different moral perspectives upon Conclusion C.

3.1 *Agent-centered and virtue theories*. Many traditional and contemporary moral theories focus on the morality of people and their lives. Virtue ethics is the oldest and in many ways the most diverse of the great ethical traditions that define Western moral philosophy, since unlike other approaches there is no one dominant conception about the analytical basis for what would constitute a good person or a good life. Nevertheless, many different theories in virtue ethics align in their accounts of particular character traits, or virtues, such as temperance, courage, justice, and wisdom, which have long been held to form an essential part of good moral character. In this subsection, we consider moral theories that fall into this category broadly construed, including all theories that consider morality as a response to the question of how one ought to live, rather than to the more specific question of what rules one ought to follow, or what outcomes one ought to bring about.

Let's begin with the observation that, as Eliezer Yudkowsky (2008) writes, "people who would never dream of hurting a child hear of an existential risk, and say, 'Well, maybe the human species doesn't really deserve to survive.'" Along these lines, surveys report that religious people, in particular, "overwhelmingly do not believe that humans will become extinct." This suggests that some individuals seem to maintain that avoiding human extinction isn't necessarily constitutive of a virtuous character; it isn't part of the "good life" because it isn't a proper concern for human beings—at most, the continued survival of humanity is God's business. Yet even this highly personalistic and intuitive sense of what constitutes a "good person" is compatible with arguments that compellingly support the proposition that we

---

[21] Tonn is obviously considering the specific case of humanity remaining on Earth rather than colonizing the universe.

ought to be seriously concerned about the fate of our species. For example, consider Bostrom's (2013) statement that

> we might also consider the issue from a less theoretical standpoint and try to form an evaluation instead by considering analogous cases about which we have definite moral intuitions. Thus, for example, if we feel confident that committing a small genocide is wrong, and that committing a large genocide is no less wrong, we might conjecture that committing omnicide is also wrong. And if we believe we have some moral reason to prevent natural catastrophes that would kill a small number of people, and a stronger moral reason to prevent natural catastrophes that would kill a larger number of people, we might conjecture that we have an even stronger moral reason to prevent catastrophes that would kill the entire human population.

Yet most agent-centered and virtue-ethical theories can say much more about the wrongness of human extinction than this. For instance, a virtue ethicist could argue that causing human annihilation would constitute evidence of a (very) bad moral character because it would, at the very least, evince some combination of imprudence, cowardice, folly, or injustice. On the flip side, pursuing actions that ensure our continued survival could be seen as manifesting virtues like compassion, courage, and love. For example, consider Philippa Foot's (2002) contention that wisdom is not merely an intellectual but a (two-part) moral virtue as well. Insofar as this is correct, it is difficult to see how the sentence "Jennifer is a paragon of wisdom who doesn't much care whether humanity survives another 100 years" doesn't engender a contradiction of some sort. Indeed, Max Tegmark (2016), following Sagan and others, argues that contemporary humanity is in a life-and-death "race between the growing power of technology and the growing wisdom with which we manage it." One could thus rephrase this as follows: "Humanity finds itself in the midst of an existential race between the growing power of technology and the *ethical virtuousness* of the persons and institutions that wield this awesome power."

Another argument for why good people should care about human extinction stems from the connection between (a) the extent to which our lives are "value-laden," i.e., "structured by wholehearted engagement in valuable activities," and (b) humanity existing far into the future under reasonably good conditions. Consider Samuel Scheffler's (2018) assertion that

> our capacity to find value in our activities here and now is more dependent *than we realize* on the implicit assumption that human life will continue long after we ourselves have died. Many of the activities that we now find it worthwhile to engage in would lose much of their point, and would seem to us much less valuable, if we thought that human life was about to come to an end (italics added).

Scheffler illustrates this with a vignette drawn from P.D. James's novel *The Children of Men*: Imagine that the entire human population were to become infertile for some unknown reason, but that this condition has no effect on the lifespans of those already living. According to Scheffler (2018), "most of us would find the prospect of humanity's imminent extinction unbearably depressing." One reason might concern the self-interested (or selfish) phenomenon of vicarious immortality, and another the fact that many activities that virtuous people engage in—from cancer research to childhood education to designing spacecraft to maintaining archives of historical documents—will partially, primarily, or entirely benefit future generations. But there may be another, more fundamental reason that stems from the idea that people engage in such activities *in the first place* because future generations *already matter* to them, i.e., if future generations didn't matter to them, then such activities wouldn't have the same gravitational pull to begin with. In Scheffler's words, referring to future generations, "we have an interest in their survival in part because they matter to us; they do not matter to us solely because we have an interest in their survival" (Scheffler 2018). Thus, many people embody and are driven by a "love of humanity," where this love, often unrecognized as such, is *revealed by the fact* that most of us would become gloomy and lassitudinous if our collective end were nigh. To quote Scheffler (2018) once more,

> suppose we agree that love is an evaluative attitude that represents its objects in a favorable light and includes a strong concern for their flourishing. And suppose we agree that to love a person is,

> among other things, to be vulnerable to a wide range of context-dependent emotions, including, most notably, distress if the person is harmed or damaged or destroyed. So too, *mutatis mutandis*, when what we love is not a person but something else. Then we may add that, in reacting as we do to the prospect of humanity's imminent disappearance, what we reveal is our love of humanity.

Thus, it is not "true that future generations are nothing more to us than potential beneficiaries or victims of our actions," as some utilitarians would argue. Rather, our descendants are "more thoroughly implicated in the structures of value that we rely on in our own lives … than we normally realize"; i.e., justified belief that the "succession of cohorts," as Scheffler puts it, will continue for an indefinitely long time under conditions that are conducive to flourishing is crucial for present humans to foster value-laden existences (Scheffler 2018).[22,23]

These claims gesture at yet another argument for valuing the continued existence of humanity (see Bell 1993). The idea goes like this: If you have children and your children have children, in perpetuity, and if you care about your children's happiness, then you should care about the happiness of your children's children, since your children's happiness will be bound up with your children's children's happiness, just as your happiness is bound up with your children's happiness. This logic applies equally to each new generation, meaning that if you care about your children's happiness, then you should also care about the well-being of your great-great-great-great-great-great-(etc.)-grandchildren. It follows that, since a human extinction event would very likely (in most cases) cause immense amounts of bodily and psychological harm, thus severely and negatively impacting individual well-being, one should work not only to ensure that a human extinction event doesn't occur within your children's lifetime, but within you children's children's lifetime, and so on. Put differently, there is an *unbroken chain of caring* that extends into the future forever and this seems to imply a duty to reduce the probability of extinction however possible. This can, in fact, be generalized beyond one's own genealogical lineage: Even childless individuals might care about, for example, their friend's children, thus leading to the same conclusion.

3.2 *Kantian and deontological theories*. Deontological theories are those which conceptualize morality in nomological terms, and rightness or wrongness as our duty to follow the moral law. Historically, deontology emerged from the philosophical work of Immanuel Kant, who argued that any universal law for rational beings, i.e., any consistent conception of morality, must be founded solely on the nature of rationality itself and should not refer to mere facts about the (nominal) world or our relation to it. More recent Kantian and deontological approaches, however, have greatly relaxed such restrictions and now believe that it is important to consider the consequences of actions as an integral part of morality. One aspect that deontological (and contractualist) theories share is that the subject matter of moral deliberations should be one's "maxim" of action, that is the rule or principle according to which one makes choices and decisions about how to act, rather than either the individual action or one's entire life.

Kant himself claims that we have a "perfect duty" to refrain from committing murder or suicide, since any maxim that might lead us to act in this way would fail to treat the rational will, either of ourselves or another, as an end in itself, and would thus contradict the notion of morality as a universal law for rational beings. Since causing human extinction would entail both murder and suicide, it seems to follow that we also have a perfect duty not to cause human extinction. Furthermore, Kant also claims that we

---

[22] Incidentally, Cheryl Misak (2016) writes that, with respect to "the structure of value, … as Scheffler (2013) argues, if we knew that human beings would become extinct once everyone currently alive died a natural death, then our own lives, contributions, and practices would diminish in value. While Scheffler is not writing about pragmatism, his insight is a *fundamental insight* of that tradition. The concepts of knowledge, rationality, and value make sense only within ongoing (although we do not have to believe them infinitely ongoing) practices of inquiring, justifying, acting, and living" (italics added).

[23] Here one might think that the concern for future generations is reducible to a concern for the perpetuation of one's own genealogical lineage, since the succession of generations is of course comprised of the succession of parents, children, grandchildren, and so on. But Scheffler (2018) maintains that this is wrong, writing that "the concern that the chain of human generations should extend into the future is not a concern that some particular line of individual descent should persist. … Dismay at the prospect of humanity's imminent extinction would not be limited to those who have or expect to have children." For Scheffler, the "love of humanity" is entirely distinct from self-interest, traces, or the motive of sympathy in utilitarian moral psychology.

have an imperfect duty to help others achieve their goals, since it is in our rational interest that everyone helps one another, rather than the opposite, given that we ourselves will often be in need of others help to achieve our own goals. Thus, it seems to follow that we at least have an imperfect duty to work to avoid human extinction, insofar as an extinction event would frustrate the achievement of such goals. A Kantian could also perhaps propose that, since humans are the only entities in the universe that are "ends in themselves," a universe without humans would be a universe without value, and a universe without value would be undesirable.[24] But even if Kantians did not want to go this far, a weaker version of the argument would point out that humanity clearly is something that ought to be valued as an end in itself—or, as Frick (2017) put it above, final value—and that respecting this value, as Kantian morality requires, implies working to maintain the existence of our species in the universe. Kant himself never made any such claims, although he does explicitly discuss the long-term future of humanity on a number of occasions, arguing that we are progressing toward ever-greater states of perfection and that past generations have sacrificed for future ones.[25] If we contemplated acting on a maxim that would allow us not to work in the interests of future generations, then it would seem that this would violate our own reliance on the work of those in the past who have benefited us.

Others have argued that Kant's notion of morality should be extended in ways that would allow a deontologist to give additional arguments in favour of prioritizing human survival. For instance, Parfit has criticized Kant's position for being overly dogmatic in its refusal to consider conditional facts about the world in assessing maxims of action; e.g., in stipulating that it is never acceptable to lie no matter what the consequences. Parfit responds to this by saying that all rational beings would agree that certain circumstances should permit lying and that the test of a moral maxim should be whether or not it would actually be rationally acceptable to everyone, rather than whether it would be theoretically consistent with a particular notion of rationality (Parfit 2011). As such, he suggests, Kantian morality should require us to do what is *optimific*, in the sense of bringing about the generally best outcome, at least so long as this is not something to which others might reasonably reject. Parfit at least clearly believes that prioritizing the avoidance of human extinction and promoting the most valuable future for humanity, in both quantitative and qualitative terms, would form a large part of what is optimific, and that this is something that, upon reflection, everyone ought to rationally accept as a key tenet of morality.

Alongside these Kantian arguments there are also several deontological arguments for procreation that suggest, as a corollary, that we should actively strive to avoid human extinction. For example, some philosophers contend that people are "[morally indebted as citizens] to the society that provided one's own education, one's opportunity to make a living, and one's civil, political, and legal rights" and that this debt can be paid off by "producing children for the benefit of society" (see Overall 2012). Others maintain "that there is a duty to perpetuate one's name, one's genetic line, and one's family property" and that this similarly entails a version of normative pro-natalism. While normative pro-natalism itself, from this deontological perspective, only implies that human extinction would be bad if it were the result of a failure to follow through on our procreational duties, there is another interpretation of both arguments that suggests a stronger conclusion. Consider first that the spirit of "giving back" to society by creating new people is based on a concern for society's flourishing. What sense would it make, though, to care about society's flourishing but not its survival? As Scheffler (2007) suggests above: What sense could it make to say, "I want X to flourish but care not whether X persists in the future"? Furthermore, Christine Overall (2012) observes that many people who are inclined to accept the second argument also tend to see procreation as a mechanism for achieving vicarious immortality. If this is true, it follows that, since vicarious immortality is conditional upon the "immortality" of our lineage, such individuals should also care about human extinction. Thus, insofar as one accepts these pro-natalist deontological arguments, one should view human extinction as bad and, therefore, important to avoid.

---

[24] Thanks to Adam Cureton (personal communication) for pointing this out to us.

[25] For example, Kant writes that "the history of mankind can be seen, in the large, as the realization of Nature's secret plan to bring forth a perfectly constituted state as the only condition in which the capacities of mankind can be fully developed, and also bring forth that external relation among states which is perfectly adequate to this end," and that "earlier generations appear to carry through their toilsome labor only for the sake of the later, to prepare for them a foundation on which the later generations could erect the higher edifice which was Nature's goal" (Kant 1784).

3.3 *Contractualist theories*. Contractualist moral theories are, like deontological theories, concerned about the moral law and our duty to obey it; indeed, until relatively recently the two were considered as falling into the same family, and many continue to view them as such. However, contractualist moral theories differ from purely deontological theories in their conception of where the moral law comes from. For classical deontologists, the moral law is a product of rationality itself and would exist even in a universe with only one moral agent. Contractualists, on the other hand, see the moral law as stemming from the existence of multiple moral agents and our need to respect one another as equals. The test for a moral maxim is thus not whether it can be rationally accepted by everyone, but rather whether it could be "reasonably rejected" by anyone, i.e.., whether we could justify acting on that maxim to everyone who might be affected by the action.

According to Elizabeth Finneron-Burns (2017), Scanlonian contractualism views human extinction as bad for two, and only two, reasons: (a) "It would cause existing people to suffer pain or death," and (b) "it would involve various psychological traumas." This suggests that one should work to avoid human extinction that is either directly caused by some action or indirectly allowed by some inaction; indeed, Finneron-Burns does not distinguish between the two, although virtue ethicists and deontologists almost universally do, seeing only the former as morally relevant. However, Finneron-Burns is explicit that the state or condition of being extinct—(S)—whereby the extinction event "would prevent millions of people from being born" and "mean the loss of rational life an civilization," is not bad from this perspective because it represents a mere absence, which is not a burden on anyone and hence could not be the source of a maxim's reasonable rejection. Along similar lines, Rahul Kumar (2009) argues that, with respect to this form of contractualism,

> there is one important question regarding future generations that might be thought to appeal to moral norms that fall outside that aspect of morality which contractualism aims to illumine. [Scanlon's view] appears to say nothing about the idea that there is something morally objectionable about doing what will ensure that no one is living in the further future. It is an open question as to whether anything at all can be said to better illumine this idea, to the extent it is defensible, by appeal to ideas implicit in the contractualist framework.

However, Kumar believes that though we cannot be said to harm future people by causing them not to exist, he doesn't share Finneron-Burns' view that a contractualist cannot object to causing or allowing human extinction because of the lives that would never be lived, as well as because of the direct harm we might do to the living. There are clearly cases, he argues, in which we act wrongly by failing to take a person's interests into account even though we do not, in fact, harm or burden them in any way. One example in which we can be said to do this is when we act negligently, or even viciously, towards a person but fail to do any harm through sheer dumb luck. If this is possible, the argument goes, then why is it not equally possible for us to be failing to take somebody's interests into account in cases in which that person is not burdened because they do not actually exist? (Kumar 2018)

Others have argued that non-existence can be treated as a burden for individuals who would otherwise live a good life had they existed, or at least that these two things can be treated analogously in certain cases (Author #2). Such views are often rejected by contractualists on the grounds that they appear to yield counterintuitive implications, such as that maximizing one's progeny is a strong moral duty for all people. However, one could formulate a duty to promote the well-being of future people that is sufficiently weak to avoid these implications while still being strong enough to provide a meaningful duty to promote the survival of humanity.

Finally, even if we accept that we have no contractualist obligations to ensure the survival of our species it may be that we have obligations to the past. As explored above, one can see human history as a large-scale cooperative endeavor, and it is no doubt true that many people in the past have made sacrifices (a) for the specific benefit of future generations, and (b) to achieve vicarious immortality. These people may well rationally object to the present generation not making reasonable sacrifices to ensure that the continued survival of humanity. They might do so on the grounds that human extinction would actively harm such people postmortem, along the lines of Bennett's proposal in subsection 2.5.3, or that while we make them no worse off, we nevertheless make their sacrifices un-worthwhile, and hence unfair (Kaczmarek 2018). This notion that we owe it to past generations to promote the interests of future people is, to

be sure, now well-developed in Western philosophical traditions, it has clear parallels with many traditional, communitarian, approaches to intergenerational ethics including those explored in contemporary African philosophy (Behrens 2012).

One might here wonder whether the political (rather than moral) contractualism of John Rawls (1971, 2001) has something more to say about the future. After all, Rawls was among the very first philosophers to seriously consider the topic of what obligations we have to future generations. Indeed, he proposes a "just savings principle" within his sufficientarianist framework, but this involves placing contractors in the "original position" behind a "veil of ignorance" about *which* generation they exist in. Since one cannot exist in non-existent generations, Rawls's conceptual exercise is mute about the potential badness of human extinction.

3.4 *Consequentialist theories*. Finally, let's consider consequentialist moral theories, which offer the standard framework within which the wrongness of extinction tends to be evaluated, and which many people believe offers the strongest reasons for prioritizing the survival of humanity. This approach views moral decision-making as a matter of selecting between possible outcomes in order to produce the best results. Traditionally, this selection is made when we are deciding between different possible acts that we might perform (act consequentialism), although it can also be undertaken when selecting between rules or maxims of action, between personal character traits we might cultivate, or any other feature over which we have control. In each case, the consequentialist enjoins us to consider what outcomes may result as a consequence of our choices, and to choose in whatever way is likely to produce the best outcome according to some axiological procedure.

This being said, let's begin by noting that some consequentialist positions positively prescribe our extinction. For example, David Benatar (2006) argues that the best outcome for our lineage is what he calls a "dying-extinction," whereby the global population dwindles to zero as a result of the voluntary cessation of all procreative activities. Yet, as one of us has elsewhere shown, Benatar's harm-benefit asymmetry does not actually entail that humanity must go extinct; i.e., a universally implemented global anti-natalist policy does not necessitate human extinction, since current or future people could use radical life-extension technologies to live indefinitely long lives (see author). Another example comes from negative utilitarianism, which, on a sufficiently "strong" or "radical" interpretation, implies that we have a moral duty to annihilate all sentient life in the universe. Other kinds of consequentialism, such as so-called "person-affecting utilitarianism," hold that the value of our action's consequences depend solely upon how these actions will affect a subgroup of people. This might be the group who are alive at the point at which we chose our actions, the point at which we can be said to have completed them, or the group whose existence depends in no way upon the choices of our actions (and who will hence exist whatever we decide to do). Such theories will generally exclude from the evaluation of outcomes the value of all those lives that will only exist if we either do not bring about human extinction, or successfully prevent it from happening, although they will still see the actual effects of human extinction on people's lives as bad.

Other consequentialist moral theories care about the lives of possible future people but reject the Total and Simple Views; such theories instead care mostly about the quality of people's lives. For example, the "Average View" asserts that only the average welfare level across all existing people matters morally, which implies that the total number of people who exist at a given moment or across time is immaterial to the morally relevant outcome. Versions of the Average View, though, may still accept that avoiding human extinction should be a moral priority, since, as argued above, the average welfare level of past humans, or indeed past sentient life, could be very low, whereas the amount of welfare in the future could be extremely high. Thus, there is a real chance of altering the balance of happiness and suffering in the universe, which could increase the overall diachronic welfare average. Alternatively, some consequentialist moral theories care predominantly about the quality of future lives because they are "Perfectionist" in the sense that they give far greater weight to qualitative considerations. As Parfit (1984) writes,

> these people believe that there is little value in the mere sum of happiness. For these people, what matters are what Sidgwick called the "ideal goods"—the Sciences, the Arts, and moral progress, or the continued advance towards a wholly just world-wide community. The destruction of mankind would prevent further achievements of these three kinds. This would be extremely bad

> because what matters most would be the highest achievements of these kinds, and these highest achievements would come in future centuries.

Finally, there are consequentialist moral theories that accept the Total, or at least the Simple, View and would thus see the loss of both an astronomical amount of quantitative and qualitative value in the future as contributing (significantly) to the tragedy of human extinction. There are strong reasons for endorsing such views. For example, according to Total and "Critical Level" Utilitarianism as well as Parfit's (2016) "Imprecise Lexical View," it is absolutely crucial to ensure that humanity does not go extinct, since an extinction event would permanently preclude the realization of such value. Unless there are extraterrestrials that could fill the universe with human-recognizable value in our absence, we ought to *hyper*-prioritize the avoidance of our collective demise. Indeed, if one combines the unique hazards hypothesis with the astronomical value thesis, one gets what could be called the "bottleneck hypothesis," which states the following:

> *Bottleneck hypothesis*: Although humanity finds itself in a period of historically unprecedented dangers to our long-term survival and prosperity, the *actual* value of the future could be astronomically large, if we play our cards right.

Incidentally, this comports with the "maxipok rule" proposed by Nick Bostrom (2013), according to which "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." The bottleneck hypothesis, though, adds a sense of *urgency* to this situation by underlining the apparent fact that, as the late Stephen Hawking (2016) declares, "this is the most dangerous time for our planet."[26]

Section 4. Weighing Future Prospects Under Normative Uncertainty

Having now explored the most prominent moral perspectives, let's turn to the remaining cluster of arguments for Conclusion C. To begin, consider the simple observation that human extinction would be irreversible; thus, it is possible to talk about our extinction occurring tomorrow but *never* about it having occurred yesterday.[27] It follows that if one were to either cause or allow human extinction to occur, one should be *absolutely certain* that this is the correct choice. But since one could never satisfy this epistemic requirement, then one should never cause or allow human extinction to occur.

Furthermore, humanity appears to have made considerable strides in ethics, both in terms of its philosophical study and its practical implications, although we are a very long way from perfecting either. Moral theory in its secular form has often progressed in fits and starts, with long periods of retrenchment during which people allowed the "wisdom of the ancients" to stand in the place of critical reflection. Hence, practices and institutions such as slavery, misogyny, deliberate obfuscation of the truth, and the institutionalization of sadistic impulses had gone unchecked, and even endorsed, by the supposed highest arbiters of our species' ethical judgement. This has been the case for millennia, of course, before finally coming under sustained attack only within the last 300 years or so, often hand in hand with the decline of religious dogmatism as a source of public authority. Yet we doubt that even the most optimistic people would claim either (a) that these advances have yet been secured for our species as a whole, or (b) that we have reached a state of perfection in either our understanding or our implementation of ethics. Of course, some people find it hard to live with the idea of their own immorality and try to customize a moral theory that says that they are already doing everything they should. For us, however, this is the exact equivalent of people who, unable to deal with how much they currently think they know will be proven wrong, and how much that they have never even guessed at will be discovered, try to believe that they know all there

---

[26] The one possible exception, of course, is the catastrophic eruption of the Toba supervolcano some 75,000 years ago, which genetic evidence indicates may have reduced the global population to somewhere in the region of 1,000 to 2,000 individuals.

[27] In an information-theoretic sense (see author).

is to know already. We live on the shore of a vast sea of ethics that we have barely stuck our toes into. And yet we are amongst the most ethical beings who ever lived.

That this is so should be self-evident once one considers all those, human and animal, who live entirely without a concept of morality or a way of thinking about what is best. While these beings don't always do bad things, to the extent that they do anything good it is purely on instinct. Herbivores, for instance, have much more compassionate diets than most humans (even human vegetarians). However, they have no concept of compassion. Carnivores are surely neither better nor worse than them. Human beings who can eat meat and would like to do so, but who choose not to, are close to a miracle and this should not be overlooked. One should also consider all those who, whilst they have a concept of morality use it merely to unreflectively implement the will of their supposed superiors, be they kings or gods, and who choose to be ignorant of the real nature of these principles. Finally, one should consider all those who do engage in some sort of reflection on the kind of person they want to be, but finding it too hard simply give up and do what they were going to do anyway.

Anyone who doesn't fall into one of these categories, who has a concept of morality, engaged in some sort of reflection on that concept, and made a positive choice that it really does provide good advice about what the right thing to do is and who has even simply continued to try and implement this advice in their life, even if they often fail, is one of the most moral beings who has ever lived. Very many humans manage to meet this bar for a sustained period of time, and we doubt whether any non-human animals even approach it (although we may be mistaken about that), and yet there is clearly so much more that could be done. Given this lack of moral progress, and our potential as a species to do so much more, it would surely be foolish in the extreme to think that we know enough already to evaluate the tragedy of human extinction, or to give up on our species as irredeemable and worthy of annihilation.

A related idea has been given the titles of "normative uncertainty" and "moral uncertainty." As Bostrom (2013) writes,

> our present understanding of axiology might well be confused. We may not now know—at least not in concrete detail—what outcomes would count as a big win for humanity; we might not even yet be able to imagine the best ends of our journey. If we are indeed profoundly uncertain about our ultimate aims, then we should recognise that there is a great option value in preserving—and ideally improving—our ability to recognise value and to steer the future accordingly. Ensuring that there will be a future version of humanity with great powers and a propensity to use them wisely is plausibly the best way available to us to increase the probability that the future will contain a lot of value. To do this, we must prevent any existential catastrophe.

William MacAskill (2014) similarly contends that, "in general, when one has the choice between two options, one of which is irreversible, and one expects to make moral progress, then option value gives one *additional* reason in favour of choosing the reversible option." Notice that, when the conditionals of being "profoundly uncertain about our ultimate aims" and "expecting moral progress in the future" are satisfied, this provides a categorical, or non-negotiable, reason for avoiding extinction. For example, consider the following scenarios: In scenario A, the future possibility of discovering a very strong argument for avoiding human extinction—say, an argument as compelling as *any* that philosophers have ever devised, one to which even the most intransigent moral skeptics would assent—constitutes a reason for ensuring our survival. In scenario B, even if future philosophers were to discover a very strong argument for either causing or allowing human extinction—say, one with equally powerful effects on the intellectually honest mind—we would *still* have reason to continue surviving because future research could discover a critical flaw in this argument, an even more powerful counterargument, and so on. The point here overlaps, at least to some extent, with what Tonn (2009) calls the "maintaining options criterion," which "entails gifting to our posterity future worlds that are as free of man-made constraints as possible." Since the ultimate constraint—or so one might argue on certain population-ethical assumptions—would be non-existence, then we should always strive to avoid our extinction.

Another argument could be made from a more impersonal perspective. Consider that R. Jay Wallace (unpublished) argues that the reason for worrying about the continued survival of humanity is that this is necessary for the continued survival of our *values*. That is, our values are what ultimately matter, not the ongoing succession of generations—contra Scheffler (2018). Now ponder John Barrow and Frank

Tipler's (1986) eschatological claim that although "our species is doomed, our civilization and indeed the values we care about may not be." The reason is that,

> from the behavioural point of view, intelligent *machines* can be regarded as people. These machines may be our ultimate heirs, our ultimate descendants, because under certain circumstances they could survive forever the extreme conditions near the Final State [of the universe]. Our civilization may be continued indefinitely by them, and the values of humankind may thus be transmitted to an arbitrarily distant futurity.[28]

Thus, if future people are unable to leap from one realm to another in the multiverse, it could still be possible—or so Tipler and Wheeler argue—to indefinitely sustain the presence of those values that we care about. This hints at the possibility of a *civilizational trace* that could outlast carbon-based biological life in our entropy-governed universe. The arguments from irreversibility, normative uncertainty, and keeping our options open thus imply that existing is better than not existing, at least so long as we are sure that there will be no other species to survive us, and consequently they provide yet another, even stronger, reason for accepting Conclusion C.

This being said, there is yet another argument for the preservation of our lineage that begins from the premise that we are almost certainly not alone in the universe. If this is the case, and if the development of civilization by intelligent beings follows a relatively linear path through something like an industrial phase to a postindustrial phase to a spacefaring phase that might be enabled by von Neumann probes that are launched ahead of colonization expeditions to collect negentropic resources and terraform exoplanets, then we might expect the universe to eventually become filled by extraterrestrial intelligences if we don't do this first. It follows that *if* such intelligences were to hold values that are antithetical to our own, it might be important (to us) to preempt their expansion into the cosmos to obviate a future in which the universe becomes saturated with astronomical amounts of disvalue (again, from our human axiological perspective). To be clear, this idea—call it the "argument from cosmic preemption"—is highly speculative. Nonetheless, it has been gestured at by numerous futurists who are concerned about ensuring that the future, on cosmological timescales, turns out well. It would be unfortunate if, say, our species fell victim to death and another species with violent, intolerant, bellicose, or "evil" tendencies were to become the dominant lifeform in our future light cone.

Section 5: Conclusion

The present paper has, admittedly, traded depth for breadth. Yet we believe that this is warranted by the facts that (a) no comprehensive survey of arguments for Conclusion C has yet been published, and (b) there is a growing need, in our view, for "big-picture" perspectives on how the great experiment called "civilization" could, or will likely, turn out. The central aim of this paper was, at minimum, to demonstrate that there are multiple independent lines of reasoning that all converge upon the basic idea that human extinction would be bad, and thus to argue that we ought to prioritize its avoidance; a subsidiary point was to affirm that one need not be a utilitarian, or even a consequentialist, to care about the long-term sustainability of our evolutionary lineage. In other words, there is a larger "meta-argument" whose consequent is Conclusion C and whose antecedent consists of a disjunction of distinct ideas and multiple tragic features of human extinction. This is epistemologically notable because disjunctive propositions are (under certain conditions that here apply[29]) necessarily more probable than single propositions, which are necessarily more probable than conjunctive propositions. The result is a powerful case that, for numerous moral and non-moral reasons, humanity should take the possibility of annihilation far more seriously than it has and currently is taking it.

This being said, we should not end this paper without examining a few reasons why human extinction might be good, that is, from the particular perspectives outlined above (rather than from the perspective of explicitly pro-extinction value systems, such as Benatarian anti-natalism). First, there could be

---

[28] I am indebted to Trisel 2004 for reminding me of this passage.

[29] E.g., few of these disjuncts are in tension with each other or are mutually exclusive.

states of existence that are univocally worse than (S), however bad (S) may be; call these "hyper-existential risks." It follows that a genuinely altruistic, anti-extinction agent might wish to take action to annihilate humanity if she were to determine that an imminent hyper-existential risk was more or less inevitable. On a similar note, one might adopt a "suffering-focused ethics" according to which the worst-case outcome for life in the universe would be the instantiation of a "suffering risk" (or "s-risk"), which is defined as "events that would bring about suffering on an astronomical scale, vastly exceeding all suffering that has existed on Earth so far" (Althaus and Gloor 2018). As with before, someone with extremely strong evidence that a suffering risk is about to occur might, out of moral feelings of empathy and sympathetic concern take it upon herself to bring about our extinction. Bostrom (2013) himself acknowledges this complication, writing that

> it is on no account a conceptual truth that existential catastrophes are bad or that reducing existential risk is right. There are possible situations in which the occurrence of one type of existential catastrophe is beneficial—for instance, because it preempts another type of existential catastrophe that would otherwise certainly have occurred and that would have been worse.

From a consequentialist and virtue ethics perspective, then, there may be circumstances in which causing or allowing human extinction would elicit moral approbation. However, since refraining from murder and suicide are perfect Kantian duties, there might never be an occasion on which we had a deontological duty to euthanize our evolutionary lineage to prevent it from experiencing worse-than-extinction or astronomical amounts of suffering.

Finally, it may be worth mentioning the phenomenon of "phyletic extinction," also called "pseudo-extinction," which occurs when one species morphs into another through anagenetic evolution. In fact, we gestured at this idea previously when discussing the possibility that we become a radically enhanced posthuman species, a process that could entail the complete disappearance of *Homo sapiens*. If the resulting species (or group of species) were genuinely "better" in some ethically robust sense, then this form of absence—being replaced by presence—would not be bad at all. One of us has elsewhere suggested that, if Persson and Savulescu (2012) are correct about *Homo sapiens* being "unfit for the future," then it may be the case that the only way for "us" to survive is to go extinct, in the phyletic sense above. But, of course, to go extinct in this manner requires that we continue to survive until sufficiently powerful person-engineering technologies become available.

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-suffering-a-neglected-priority/.

Barrow, John, and Frank Tipler. 1986. *The Anthropic Cosmological Principle*. Oxford: Oxford University Press.

Behrens, Keven, 2012. Moral obligations towards future generations in African thought. Journal of Global Ethics, 8(2-3), pp.179-191.

Bell, Wendell. 1993. Why Should We Care About Future Generations? *The Years Ahead*. Bethesda, MD: World Future Society.

Benatar, David. 2006. *Better Never To Have Been: The Harm of Coming Into Existence*. Oxford: Oxford University Press.

Bostrom, Nick. 2003a. Are You Living in a Computer Simulation? *Philosophical Quarterly*. 53(211): 243-255.

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

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. 2008. Why I Want to Be a Posthuman When I Grow Up. In Bert Gordijn and Ruth Chadwick (eds.), *Medical Enhancement and Posthumanity*. Berlin: Springer.

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

Bratsberg, Bernt, and Ole Rogeberg. 2018. Flynn Effect and Its Reversal Are Both Environmentally Caused. *Proceedings of the National Academy of Sciences*. Epub ahead of print.

Burke, Edmund. 1790. *Reflections on the Revolution in France*. In *The Works of Edmund Burke* (1905).

Chalmers, David. 2005. The Matrix as Metaphysics. In Christopher Grau (ed.), *Philosophers Explore the Matrix*. Oxford: Oxford University Press.

Ćirković, Milan. Cosmological Forecast and its Practical Significance. *Journal of Evolution and Technology.* https://www.jetpress.org/volume12/CosmologicalForecast.pdf.

Collings, David. 2014. Stolen Future, Broken Present: The Human Significance of Climate Change. Ann Arbor, MI: Open Humanities Press.

Finneron-Burns, Elizabeth. 2017. What's wrong with human extinction? *Canadian Journal of Philosophy.* 47(2-3): 327-343.

Foot, Philippa. 2002. *Virtues and Vices and Other Essays in Moral Philosophy.* Second edition (first edition 1978). Oxford: Oxford University Press.

Frick, Johann. 2017. On the Survival of Humanity. *Canadian Journal of Philosophy*. 47(2-3): 344-367.

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

Hawking, Stephen. 2016. This is the Most Dangerous Time for Our Planet. *Guardian*. www.theguardian.-com/commentisfree/2016/dec/01/stephen-hawking-dangerous-time-planet-inequality.

Jacquet, Jennifer. 2017. The Anthropocene. *The Edge*. www.edge.org/response-detail/27096.

Kaczmarek, Patrick Krystof. 2018. *A Fairness-Based Astronomical Waste Argument*. Dissertation. http://theses.gla.ac.uk/8889/.

Kaku, Michio. 2005. *Parallel Worlds: A Journey Through Creation, Higher Dimensions, and the Future of the Cosmos*. New York, NY: Doubleday.

Kant, Immanuel. 1784. *Idea for a Universal History from a Cosmopolitan Point of View*. https://www.-marxists.org/reference/subject/ethics/kant/universal-history.htm.

Korsgaard, C.M., 1983. Two distinctions in goodness. *The philosophical review*, 92(2), 169-195.

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

Kumar, Rahul. 2009. Wronging Future People: A Contractualist Proposal. In Axel Gosseries and Lukas Meyer (eds.), *Intergenerational Justice*. Oxford: Oxford University Press.

Kumar, Rahul, 2018. Risking Future Generations. Ethical Theory and Moral Practice, pp.1-13.

MacAskill, Will. *Normative Uncertainty*. Dissertation. https://www.academia.edu/8473546/Normative_Uncertainty.

Misak, Cheryl. 2016. *Cambridge Pragmatism: From Peirce and James to Ramsey and Wittgenstein*. Oxford: Oxford University Press.

Nietzsche, Fredrich, 1977. The Portable Nietzsche. Penguin.

Lenman, James. 2002. On Becoming Extinct. *Pacific Philosophical Quarterly*. 83: 253-269.

Overall, Christine. 2012. *Why Have Children?: The Ethical Debate*. Cambridge, MA: The MIT Press.

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

Parfit, Derek, 2011. On what matters: volume one (Vol. 1). Oxford University Press.

Parfit, Derek, 2016. Can we avoid the repugnant conclusion?. Theoria, 82(2), pp.110-127.

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

Pearce, David. 2007. The Abolitionist Project. https://www.abolitionist.com/.

Persson, Ingmar, and Julian Savulescu. 2012 *Unfit for the Future: The Need for Moral Enhancement.* Oxford University Press, Oxford.

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

Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press. Revised edition, 1999.

Rawls, John. 2001. *Justice as Fairness: A Restatement.* E. Kelly (ed.), Cambridge, MA: Harvard University Press.

Sandberg, Anders, and Nick Bostrom. 2008. Global Catastrophic Risks Survey, Technical Report #2008-1. www.fhi.ox.ac.uk/reports/2008-1.pdf.

Sandberg, Anders, Eric Drexler, and Toby Ord. 2018. Dissolving the Fermi Paradox. arXiv. https://arxiv.org/pdf/1806.02404.pdf.

Scanlon, T. M. 1998. *What We Owe to Each Other*. London: Belknap Press of Harvard University Press.

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

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

Seachris, Joshua. 2011. Death, Futility, and the Proleptic Power of Narrative Ending. *Religious Studies*. 47(2): 141-163.

Slaughter, R. 1994. Why Should We Care About Future Generations Now? In *Why Future Generations Now?* Institute for the Integrated Study of Future Generations. Kyoto, Japan, 96–113.

Slovic, Paul. 2007. Psychic Numbing and Genocide. *Psychological Science Agenda*. http://www.apa.org/science/about/psa/2007/11/slovic.aspx.

Tegmark, Max. 2016. The Wisdom Race Is Heating Up. edge.org. www.edge.org/response-detail/ 26687.

Thompson, Janna. 2012. *Intergenerational Justice: Rights and Responsibilities in an Intergenerational Polity*. New York, NY: Routledge.

Todd, Benjamin. 2017. Why Despite Global Progress, Humanity Is Probably Facing Its Most Dangerous Time Ever. 80,000 Hours. https://80000hours.org/articles/extinction-risk/.

Tonn, Bruce. 2009. Obligations to Future Generations and Acceptable Risks of Human Extinction. *Futures*. 41: 427-435.

Tonn, Bruce, and Donald MacGregor. 2009. Are We Doomed? *Futures*. 41(10): 673-675.

Tough, Allen. 1991. *Crucial Questions About the Future*. Lanham, MD: University Press of America.

Trisel, Brooke Alan. 2004. Human Extinction and the Value of Our Efforts. *The Philosophical Forum*. XXXV(3): 371-391.

Trisel, Brooke Alan. 2016. Human Extinction, Narrative Ending, and Meaning of Life. *Journal of Philosophy of Life*. 6(1): 1-22.

Wager, W. Warren. 1983. H.G. Wells and the Genesis of Future Studies. http://www.wnrf.org/cms/hgwells.shtml.

Wallace, Jay. Unpublished. Value, Trauma, and the Future of Humanity. https://philosophy.berkeley.edu/file/1025/Value_Trauma_Future.pdf.

Ward, Peter, and Donald Brownlee. 2000. *Rare Earth: Why Complex Life is Uncommon in the Universe*. New York, NY: Copernicus Books.

Wells, H.G. 1902. *The Discovery of the Future*. https://archive.org/stream/discoveryoffutur00welliala/discoveryoffutur00welliala_djvu.txt.

Wiblin, Robert. 2017. Why the Long-Term Future of Humanity Matters More than Anything Else, and What We Should Do About It. 80,000 Hours. https://80000hours.org/podcast/episodes/why-the-long-run-future-matters-more-than-anything-else-and-what-we-should-do-about-it/.

Wilhelm, James. 2012. Janna Thompson: Intergenerational Justice. *Intergenerational Justice Review*. 1: 34-36.

Yudkowsky, Eliezer. 2008a. Cognitive Biases Potentially Affecting Judgement of Global Risks. In Nick Bostrom and Milan Ćirković (eds.), *Global Catastrophic Risks*. Oxford: Oxford University Press.