‘Humanity’: Constitution, Value and Extinction

Elizabeth Finneron-Burns

University of Western Ontario

Forthcoming in *The Monist*

ABSTRACT

When discussing the extinction of humanity, there does not seem to be any clear agreement about what ‘humanity’ really means. One aim of this paper is to show that it is a more slippery concept than it might at first seem. A second aim is to show the relationship between what constitutes or defines humanity and what gives it value. Often, whether and how we ought to prevent human extinction depends on what we take humanity to mean, which in turn depends on what we value about humans. Finally, I will offer a definition of humanity that I argue offers a plausible account of when we might say humanity has gone extinct, and reflects what we value about humans, and relatedly, what many actually care about in terms of why we should be preserved.

1. INTRODUCTION

Until quite recently, the ethics of human extinction had been largely neglected by philosophers. Whether that is because most people just think it would be obviously extremely bad (or wrong, depending on its cause), or they just don’t think much about it, I can’t be sure. However, this is happily changing, and more and more philosophers and political theorists are turning their attention to this important problem. Many philosophers take it for granted that humanity’s extinction would be bad and/or wrong and move straight to determining ways to avoid it, and/or the implications of it occurring (McKinnon 2017; Leslie 1996), whereas others, myself included, have focused more on determining if it would be bad or wrong, and if so, why (Finneron-Burns 2017; Frick 2017; Kaczmarek and Beard 2020; Palazzi 2014).

When discussing this topic, philosophers tend to use the term ‘humanity’ in many different ways and there does not seem to be any clear agreement on what this term really means. Writing in *Aeon*, Emile P. Torres (2023) acknowledges this when pointing out the (often underappreciated) complexity of the ethics of human extinction: “Consider that ‘humanity’ and ‘extinction’ can be defined in many different, equally legitimate ways.”[[1]](#footnote-1) It may seem strange to claim that there is no agreement on what ‘humanity’ is because we frequently use the term in mutually intelligible ways. We talk about the survival of humanity, prosecute people for crimes against humanity, and we exclaim ‘oh the humanity!’ when something awful happens. How could we not know what it means? One aim of this paper is to bring together some of the different ways this term has been understood in order to show that it is a more slippery concept to grasp than it might seem at first. A second aim is to show the relationship between what constitutes or defines what humanity is and what gives it value. Often whether and how we ought to prevent human extinction depends on what we take humanity to mean, which in turn depends on what we value about humans. Finally, I will offer a definition of humanity that I argue offers a plausible account of when we might say humanity has gone extinct and reflects what we value about humans and what many actually care about in terms of why we should be preserved.

1. HUMAN BEINGS COLLECTIVELY

The Cambridge English Dictionary defines ‘humanity’ as the collective of all the living human beings. This is probably also (roughly) the first definition a random person would give if asked. The scientific definition of a ‘human’ is any member of the genus *homo* including, but not limited to, our own species *homo sapiens.* However, since all other member species of the genus are long extinct (for example, *homo habilis* and *homo erectus*), I will use ‘humans’ to mean the members of the species *homo sapiens.* When scientists (usually retrospectively) declare the extinction of a species, they mean that the last individual of that species has died. So, on this definition, humanity will have gone extinct when the last *homo sapiens* specimen has died, just as the dodo species became extinct when the last dodo died.

 This definition has the benefit of simplicity and makes it definitionally easy to determine when humanity has gone extinct. However, despite the simplicity of this definition, there are several ways in which it does not really capture what philosophers seem to mean when they use the term, which, as I will argue, is at least partly because many seem to fold in (not necessarily wrongly) some of the reasons to *value* humanity into the definition itself.

Furthermore, although human beings may seem to be a necessary part of humanity (more on that later), I do not think most people, when pressed, actually hold the view that humanity is simply just all living humans. Rather, ‘humanity’ almost seems like some corporate entity, independent of its constituent parts. This may sound strange so let me show you some examples of what I mean.

One example is from the last few paragraphs of Johann Frick’s paper “On the Survival of Humanity.” At the end he opens up the question of what the ‘survival of humanity’ actually means, and while he doesn’t provide a positive answer, he does offer a scenario where humans continue to exist but as discrete generations, one after the other, with no knowledge of the previous one ever having existed—starting from scratch with each generation so to speak. He says that “[t]he human species survives in this scenario, but a lot of what we mean by ‘humanity’ is lost” (Frick 2017, 362-3). The fact that he distinguishes between humanity and the mere existence of human beings in the future suggests that he thinks that humanity and the members of the species *homo sapiens* are not necessarily the same thing, and that humanity is something more than just the collection of humans. In Frick’s example, it seems likely that the ‘something’ includes humans’ connections to and knowledge of past and future people and the things that they have done or will do (more on that later).

A second example is found in Toby Ord’s recent book *The Precipice*. In it, Ord writes that “[i]n ecological terms, it is not the *human* that is remarkable, but *humanity…*Each human’s ability to cooperate with the dozens of other people in their band was unique among large animals. It allowed us to form something greater than ourselves” (Ord 2020, 12). Here Ord writes as though humanity is something separate from just the collective of humans, and also that part of ‘humanity’ is human cooperation. If humans no longer cooperated, they would still exist but not as humanity. His description of it as something ‘greater than ourselves’ suggests that he sees humanity as something that continues through time, independent of and in addition to the particular people who come and go at various points in time. Humanity would continue and be the same ‘thing’ even if every single person alive (or who has ever lived) were replaced with a different person.

 Another problem with the simplistic definition is that if we accepted that humanity were just human beings collectively, then certain scenarios would not count as the extinction of humanity that arguably should do—or, at least, that people who object to human extinction would likely find just as troubling as extinction itself.

Last Man Standing

Imagine that some catastrophe occurred, reducing the human population to such a low level (perhaps just one man and one woman) that the surviving people are just about able to sustain their physical needs for survival and reproduction, but not able to rebuild the population. Human extinction has not occurred since there are still some surviving people. But is there any normatively relevant sense in which the continued existence of this very small population is significantly different from complete human extinction?

It is true that a classical total utilitarian would think that there is more value in a few lives lived than in zero lives and therefore consider the *Last Man Standing* scenario to be better than extinction, at least assuming that the remaining people still have positive welfare after such a catastrophe. But I suspect that most of us would think: what’s the point? More to the point, I doubt if people who worry about trying to reduce existential risk would be satisfied by the response: “don’t worry, a few happy people survived!” In this scenario human extinction has not occurred but it does seem that *humanity* is over.[[2]](#footnote-2) Nick Bostrom (one of the few to actually provide a definition) defines ‘humanity’ as Earth-originating intelligent life (i.e., at least for now, human beings collectively). Yet even he agrees that “a future in which a small human population ekes out a miserable existence…would constitute an existential catastrophe” (Bostrom 2013, 24). If you agree with this, then it isn’t just the presence of the species *homo sapiens* on Earth that matters, but something else—I will provide some options for what that ‘something’ could be later on.

At one point in his defence of the ‘final value’ of humanity, Johann Frick makes the interesting point that “cherishing, protecting, respecting, savouring, etc. are all modes of responding to something’s final value which do not commit one to wanting to have more instances of it” (Frick 2017, 359).[[3]](#footnote-3) His comment is in response to what he calls the:

*Argument from Additional Lives*: if the reason to avoid human extinction is to ensure more happy people are created, then there is nothing worse about many trillions of people existing all at once and then going extinct, compared to the same number of people existing but over a much longer period of time.

Frick wants to argue that the reason to prolong humanity’s existence diachronically does not also provide a reason to make humans synchronically more numerous.

Frick’s argument is really interesting on its own, but also because it brings to light a way in which the meaning of ‘humanity’ is complicated. Frick’s quote can be applied generally, but since he is trying to defend the survival of humanity specifically, let’s plug ‘humanity’ into the quote above:

Cherishing, protecting, respecting, savouring, etc. are all modes of responding to [humanity’s] final value which do not commit one to wanting to have more instances of [humanity].

Of course, while we can, at least in theory, cherish the final value of humanity, we cannot have more instances of it (humanity). We can have more instances of *humans* (i.e., more people), but not more humanities, at least as long as humanity is understood to be human beings collectively. Even if multiple sets of humans existed independently on different planets, the definition we are currently working with would mean all those groups constituted one humanity. For Frick’s response to the Argument from Additional Lives to work, he must mean the value of *humans*, not humanity. In fact, the quote works quite well that way:

cherishing, protecting, respecting, savouring, etc. are all modes of responding to [humans’] final value which do not commit one to wanting to have more instances of [humans].

The problem is that if we follow this final formulation of Frick’s claim, then respecting human value does not commit us to creating more humans. Therefore, on his view we could respect the value of humans while at the same time allow human extinction, because, of course, that is what happens if we fail to create more instances of humans. Frick has confirmed to me that he does not intend this conclusion.

I think Frick is correct that respecting something’s value and wanting to protect and prolong it does not commit us to creating more of that thing. We can clearly respect the value of the Great Sphinx of Giza in Egypt and want to ensure it continues to be around for future people to enjoy without needing or wanting to create more Sphinxes.[[4]](#footnote-4) The problem is that humans are not like the Great Sphinx because, to put the point starkly, humans die. If we do not have humans, we cannot have humanity (on our working definition).

Humanity is less like the Great Sphinx and more like a forest. Respecting the value of and wanting to prolong the existence of a forest doesn’t necessarily commit us to making more forests, but it does commit us to planting more trees in the existing forest since each tree in the forest will eventually die and if none are replaced, the forest will cease to exist.

Likewise, the way humanity continues *is* by creating more humans. If we want to prolong the ‘forest’ of humanity, then we have to ensure that there is some critical mass of humans, which does mean we have to create at least some future humans, but how exactly many is unclear, and we end up with some very abstract questions. How many trees do we need to say there is a forest? How many humans does it take to constitute humanity?

It’s possible to avoid these sorts of abstract questions (at least ‘how many humans do we need to constitute humanity?’) by seeing humanity as something separate from humans ourselves, but then we would need an account of how we can prolong humanity without creating more humans.

At the very least, Frick’s argument against the Argument for Additional Lives shows the need to clearly define what is meant by ‘humanity’. If it is humans collectively, then we can respect the value of humans without creating more of them—but this allows human extinction. If humanity means something greater than individual people, then we can respect the value of humanity without being committed to creating more humans. But then we need an account of how we can prolong humanity without creating more humans. I will provide a possible such account in Section III.

The Stupid Virus

There is a second way in which we might think humanity had ended/was extinct, despite humans continuing to exist. Think about a virus spread through the Earth, removing people’s ability to reason. Or imagine that humans become so very dependent on technology that we are robbed of our intelligence.[[5]](#footnote-5) Humans would still be *homo sapiens*, but would no longer be capable of complex language or creating art, music, scientific discoveries, etc. We could also conjure up Bostrom’s (2002) examples of ‘Crunch’ events where modern civilization collapses and is unable to recover its previous technological and cultural advances, let alone progress any further, or where dysgenic pressures are so strong that human beings evolve to be extremely numerous but unintelligent.

Has humanity survived in these sorts of cases? If we are inclined to say ‘no’, this must be because humanity is not simply the existence of human bodies, but includes at least some of the abilities and achievements that humans have and have made. It also suggests that wanting *humanity* to have a future is not the same as wanting the species *homo sapiens* to survive. I suspect that many people would insist that the survival of humanity also means the survival of the things that make humans unique, including intelligence, culture, and the capacity to progress and achieve. Without those things, humans would not really be different from any other animal species. Although there are, of course, many people who lament the extinction of *any* species, most people probably think human extinction would be uniquely bad, and, in my view, that is not because there is something inherently better about *homo sapiens*, but because of the unique attributes of our species that would be eliminated in The Stupid Virus scenario above.

 This broadly aligns with the views of a group that Derek Parfit thinks would argue for the unique badness of human extinction. He 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 (Parfit 1984, 454).

For the group to whom Parfit is referring in this passage, the reason why human extinction would be very bad is that it would prevent human achievement—whether in the arts, sciences, or justice. If that is right, then they should not consider The Stupid Virus scenario to be preferable to human extinction because in both cases what matters (human ability) is lost. Humans continue, but humanity does not.

Hell on Earth

There is also a third way in which the species *homo sapiens* could continue to exist, but where we might think humanity does not. This is a scenario in which every human were to live a wretched life. This could be for any number of reasons including climate catastrophe or a repressive and totalitarian world government as in Bostrom’s (2002) ‘Shriek’ scenario.

If we just see humanity as a collection of humans who happen to exist at any given moment, then it is easier to argue that, in this case, extinction would not only not be bad or wrong, but might actually be a *good* thing. Utilitarians and non-consequentialists can both agree that suffering is usually a bad thing. Extinction would relieve everyone of their suffering and ensure that no further people had to endure it. I suspect that most would not consider humanity to have survived in the Hell on Earth case, since the people living these wretched lives would not be enjoying the benefits of human flourishing or achievement (or, for Bostrom, the goods of the posthuman future).

This scenario also shows how the definition and value of humanity may converge or diverge. Frick argues that the value of humanity “gives us a reason to ensure, not that humanity grimly soldiers on, but that it survives in a flourishing state” (Frick 2017, 360). It’s clear that Frick finds human flourishing to be an integral part of the point, or *value* of humanity since he says that we have strong reasons to preserve human flourishing. What is less clear is whether he thinks that humanity itself would end if flourishing ceased (i.e., human flourishing is partially constitutive of humanity), or whether humanity would continue to exist but lose its value and therefore also our reason to preserve it (i.e., human flourishing is value-giving). In either case it seems as though Frick should accept that human extinction in *Hell on Earth* would not be a bad thing—though obviously in his view it would be better if there were flourishing people. If humanity itself ends with the cessation of flourishing, then in order to argue that human extinction is wrong, we need a reason to continue to create unhappy human bodies in the absence of humanity. If humanity continues and just loses its value, then to argue against extinction we need some reason other than flourishing to preserve humanity.

Some, like Ord, write as though they consider human flourishing to be a part of what humanity means, or what constitutes it: humanity’s extinction “would foreclose our future. It would destroy our potential. It would eliminate all possibilities but one: a world bereft of human flourishing” (Ord 2020, 36). I think there are two interesting points in this short passage. The first is that Ord seems to find human flourishing to be an essential part of humanity. After all, he did not say that what is left is a world bereft of *humans*, but of *human flourishing.* For Ord, one reason to avoid extinction is to avoid ridding the world of flourishing humans. This suggests that he might agree that a world in which everyone lived wretched and decidedly non-flourishing lives would not necessarily be better than human extinction. Or, at least, he might agree that human extinction and the extinction of humanity are not the same thing, implying that humanity is not simply just the species. The second is that Ord writes as though humanity is something separate from its constituent parts when he refers to ‘*our* future’ and ‘*our* potential’ (italics mine). I am currently forty years old. The average Canadian woman lives to be about 83, meaning *my* future is probably only about forty-three more years (knock on wood!). It’s unlikely that I will accomplish anything particularly impressive in those forty-three years, certainly not the types of things that Ord likely has in mind. I am probably not going to make any scientific discoveries and if anyone had heard my singing voice, they’d know I won’t be performing at the Royal Opera House any time soon. I probably won’t even come up with some groundbreaking philosophical theory. And yet, I am included in Ord’s ‘our’ future and potential. On this view, humanity is not really about the individuals who make it up, the people who live in the future and the people who make the discoveries.

If human flourishing is a necessary part of humanity, then in a *Hell on Earth* scenario humans could continue to exist, despite humanity having ended. Further, if human flourishing is what gives *value* to humanity, then we might prefer human extinction in such a case since what is valuable about humanity is already gone. On the other hand, if humanity is simply the human species, and we think that species has value independent of the aggregative value of the people who actually exist, then continuing humanity could, at least in one sense, be a good thing, despite the poor quality of life the individual humans have.

1. HUMAN ACCOMPLISHMENT, ABILITY, FLOURISHING

All this suggests that there is a possible alternative definition of humanity in which humanity is not the collection of human beings themselves, but human achievement (scientific and technological progress, arts, culture, etc.), abilities (reason and intelligence), and flourishing, including social relationships and connections. And on this account, what also makes humanity *valuable* is not the particular species that constitutes it, but what it can *do*. Frick, for example, argues that what is special about humanity is “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” (Frick 2017, 359).

It’s important to recognise that if we accept a definition where humanity is not the species, but its capabilities, then we could have humanity continue in the absence of any humans. As I have already noted, Bostrom defines humanity as “Earth-originating intelligent life rather than as the particular biologically defined species Homo sapiens” (Bostrom 2013, 20) and Ord says that he chooses to talk about ‘humanity’ as the collection of humans only because “humans are the only beings we know of that are responsive to moral reasons and moral argument—the beings who can examine the world and decide to do what is best. If we fail, that upward force, that capacity push toward what is best or what is just, will vanish from the world” (Ord 2020, 38). He goes on to say that “[i]f we somehow give rise to new kinds of moral agents in the future, the term ‘humanity’ in my definition should be taken to include them” (Ord 2020, 39).[[6]](#footnote-6)

 Think of a futuristic scenario where human beings all uploaded our minds to robots or machines:

An upload is a mind that has been transferred from a biological brain to a computer that emulates the computational processes that took place in the original biological neural network. A successful uploading process would preserve the original mind’s memories, skills, values, and consciousness (Bostrom 2002).

In Bostrom’s quote above human minds have been uploaded to computers, but we could also imagine a scenario in which artificial intelligence becomes so advanced that it achieves consciousness itself. These machines (whether human brain uploads or artificial intelligence) would be intelligent and cultured, continue to make scientific discoveries, and experience emotions—many, if not all, of the elements that make a human life flourishing. If humanity is just about intelligence/the ability to reason, cultural experiences, and scientific accomplishment, then humanity would continue despite there being no more *homo sapiens* specimens. And if an integral part of humanity is flourishing, well, there is no reason to assume that these robots are not in fact flourishing. Indeed, they may have even longer ‘lifespans’ than humans do, given they will not be susceptible to disease. Furthermore, without nerves, the uploaded minds in robot ‘bodies’ would not feel physical pain (though emotional pain would likely persist) and it’s possible that the amount of work needed to survive would be vastly decreased, freeing up much more time for robot games and leisure.[[7]](#footnote-7) If we see humanity as something separate from actually-existing humans, then humanity could continue in the absence of people since all of these things (progress, reason, flourishing) could be done by uploads or artificial intelligence.

As we know, some philosophers argue that it benefits particular people to exist with worth living lives, or that it is impersonally good for people to exist with worth living lives. As I’ve set out the case, these uploads or artificial intelligence would likely experience much better than merely worth living ‘lives’. Those who argue against human extinction on the basis that it would prevent people from existing with worth living lives, will need to explain why, if at all, human extinction in the uploads/artificial intelligence case is bad or wrong.

Some philosophers might bite the bullet and accept that there *is* nothing wrong in this case. Ord actually seems to suggest that he might take this view. He writes: “if we somehow give rise to new kinds of moral agents in the future, the term ‘humanity’ in my definition should be taken to include them” (Ord 2020, 39). Bruce Tonn (2009) suggests that future generations of artificial lifeforms could deserve the same ethical considerations as future generations of human beings. To them, the important reason to prevent extinction is to preserve the things humans have already made and/or could potentially make and/or improve (what Tonn calls Unfinished Business). Finally, Bostrom is explicit in his bite of this particular bullet:

there is no reason to suppose that the biological species concept tracks what we have reason to value. If our species were to evolve, or use technology to self-modify, to such an extent that it no longer satisfied the biological criteria for species identity (such as interbreedability) with contemporary Homo sapiens, this need not be in any sense a catastrophe. Depending on what we changed into, such a transformation might well be very desirable (Bostrom 2013, 20).

It doesn’t seem like any of these thinkers are actually concerned about humans or human extinction at all, but about what humans have or could do.[[8]](#footnote-8) If another species or machine could do the same thing, then the loss of *homo sapiens* would not be a problem in their view. In which case, given that we know that human extinction is almost certainly going to occur eventually, perhaps the focus should not be on trying to prevent or delay it by reducing human extinction risk, but on doing whatever makes it most likely that these good things (flourishing, intelligence, science, art, etc.) will continue (including developing conscious superintelligence and/or the ability to upload minds to machines) before human extinction does occur. Derek Shiller offers a version of this conclusion, arguing that since *ex hypothesi*, artificial intelligence will live significantly better lives than human beings, “we should engineer the extinction of the human race in order to route available resources to creating and sustaining them” (Shiller 2017, 396).

I suspect that many would have doubts abut this conclusion and reply that what matters is not just intelligence or culture or flourishing, but intelligence *in human bodies,* and culture *being created by human bodies,* and flourishing *enjoyed by humans*. I can certainly appreciate this thought, but a person making this kind of argument needs to explain why human bodies matter, why they are inherently valuable separate from the things those humans can do. One option is to explain what is valuable about *homo sapiens* bodies *per se* that warrants their preservation. Or, more precisely, why would a world bereft of human bodies (through human extinction) be a bad or wrong thing even though humanity itself (i.e., human capabilities/flourishing) continues? The second option is to include in the definition of ‘humanity’ that it is constituted by human bodies in addition to the other features like intelligence, accomplishment, and flourishing. Someone taking this tack, however, will still need to explain why human bodies are important. What is so special about them that they should be included in a definition of humanity and something to be specifically preserved?

 One way to explain why human beings achieving things themselves (as opposed to uploads or artificial intelligence doing these things instead) matters would be to appeal to Bernard Williams’ distinction between objective and subjective value (Williams 2006). Objectively (from the point of view of the cosmos), it wouldn’t matter whether these capacities were held by humans or robots, or any other species. However, it might matter subjectively. That is, it matters *to us* that *humans* are the species doing these things. On this view we should not include advanced artificial intelligence, aliens, or any other (intelligent or otherwise) species in what we understand to be ‘humanity’ because they are not part of our species, and the kinship we feel with our own ‘kind’ matters to us.

1. CONCLUSIONS

My conclusion is that we should define humanity in this latter way—human bodies with reason, culture, and flourishing—because it reflects what many consider to be unique and valuable about humanity and humans and produces, in my view, what is a plausible account of which of the scenarios above would constitute human extinction (lamentable or otherwise).

 In the *Last People Standing* scenario, this definition would have us conclude that humanity has indeed gone extinct since the very small population is unlikely to flourish or have the ability to continue scientific or cultural progress. Likewise, humanity will have ended in the *Stupid Virus* and *Hell on Earth* scenarios due to the loss, respectively, of intelligence and flourishing. Finally, humanity will have ended if humans are replaced with uploads or artificial intelligence, even if these machines are able to do all of the things humans can.

On the last point I diverge from philosophers like Bostrom and Ord who see these machines as part of humanity and the moral community. As a result, my claims about what humanity is and what is valuable about it suggest that existential risk prevention ought to be focused on the preservation of human beings *and* on their abilities and flourishing. This differs from some existential risk theorists, including Bostrom’s interest in existential risk prevention focusing on the pursuit of transhumanism and ‘technological maturity’. On my account it is no better for intelligent life to continue with no humans than for human extinction to occur.

 As a parting note, something interesting follows from the fact that, on the definition I have argued for, humanity is something separate from the individual humans who make it up. This is that the good of ‘humanity’ and the good of individual humans can sometimes conflict. We could increase the well-being of individual people at the cost of shortening humanity’s ‘lifespan’ and *vice versa*, we could reduce the well-being of existing people in order to ensure that humanity could continue for longer. Climate change mitigation is an example of this. If we accept that climate change is an existential risk (Harvey 2022, cf. MacAskill 2022), then avoiding extinction requires a dramatic reduction in our carbon emissions. This can be done by consuming less of everything we enjoy, including meat, travel, and modern conveniences including electronics and washing machines. This would significantly reduce the negative effects of climate change, and the risk of extinction that comes along with it, but at the cost of our own well-being. Theoretically, it could be the case that the necessary reductions in well-being in order to avoid extinction are so significant that they cause actual human suffering for the sake of the continuation of the species. Whether any value of humanity outweighs the aggregative suffering of the actual human beings who constitute it, I am not sure, but some of the longtermist media writings certainly seem to imply that it can. However, at the very least, it does raise the question of how to weigh the preservation of humanity against the well-being of actually existing people.[[9]](#footnote-9)

# References

Benatar, David (2008) *Better Never to Have Been* (Oxford University Press).

Bostrom, Nick (2013) “Existential Risk Prevention as a Global Priority” *Global Policy* 4(1).

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

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

Dietrich, Eric (2007) “After the humans are gone” *Philosophy Now* 61.

Finneron-Burns, Elizabeth (2017) “What’s Wrong with Human Extinction?” *Canadian Journal of Philosophy* 47(2-3)

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

Greaves, Hilary and William MacAskill (2021) “The case for strong longtermism”, Global Priorities Institute, University of Oxford, <https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf> (accessed July 4, 2022).

Harvey, Fiona (2022) “Humanity faces ‘collective suicide’ over climate crisis, warns UN chief” *The Guardian*, June 18, 2022.

Kaczmarek, Patrick and SJ Beard (2020) “Human Extinction and Our Obligations to the Past” *Utilitas* 32(2).

Leslie, John (1996) *The End of the World* (Routledge).

Lovelock, James (2019) *Novacene: The Coming of Age of Hyperintelligence* (MIT Press).

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

McKinnon, Catriona (2017) “Endangering Humanity: An International Crime?” *Canadian Journal of Philosophy* 47(2-3).

Ord, Toby (2020) *The Precipice* (Bloomsbury).

Palazzi, Franco (2014) “Would Human Extinction be Morally Wrong?” *Philosophia* 42(4).

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

Shiller, Derek (2017) “In defence of artificial replacement” *Bioethics* 31(5), pp. 393-399.

Tonn, Bruce (2009) “Obligations to future generations and acceptable risk of human extinction” *Futures* 41(7), pp. 427-435.

Torres, Émile P. (2023) “The ethics of human extinction” *Aeon*. February 20, 2023. <https://aeon.co/essays/what-are-the-moral-implications-of-humanity-going-extinct> (accessed May 15, 2023).

Williams, B. (2006). “ The Human Prejudice” in Bernard Williams, Philosophy as a Humanistic Discipline*,* ed., A. Moore (Princeton University Press) pp. 135– 52.

1. Torres’ article focuses on the latter question—different senses of ‘extinction’—whereas this paper focuses on the former—different senses of ‘humanity’. However, these two concepts are inextricably linked—how do we know whether humanity has gone extinct unless we know what humanity is, and what extinction is? Therefore, the two pieces necessarily cover some of the same ethical grounds. [↑](#footnote-ref-1)
2. This is, of course, vague. How many grains of sand constitute a heap? One might agree that one human is not ‘humanity’ but what about two? Three? How many humans are needed to constitute ‘humanity’ and to say that it has not gone extinct? I think the lack of a clear answer to that question is further evidence that the ‘human beings collectively’ definition of humanity lacks much useful meaning. As I will show later in the paper, it’s more likely that ‘humanity’ is defined, not by existing humans themselves, but by what humans as a species can *do* or *achieve.*  As I will also argue, however, this raises problems of its own. [↑](#footnote-ref-2)
3. According to Frick, something has ‘final value’ if “it is valuable for its own sake, independently of the instrumental value it may have for anyone” (Frick 2017, 359). [↑](#footnote-ref-3)
4. In fact, it is plausible that what makes the Great Sphinx of Giza so valuable is that there is only one. [↑](#footnote-ref-4)
5. Thank you to an anonymous reviewer for suggesting this example. [↑](#footnote-ref-5)
6. There is also the possibility that *homo sapiens* could go extinct by evolving into another *homo* species. Due to the way evolution works, this new species would likely have even higher capacities than *homo sapiens.* If this is the case, I presume most existential risk scholars would not consider *humanity* to have gone extinct nor see anything particularly bad about the extinction of *homo sapiens*. This suggests, again, that the species itself is not really what people care about preserving (beyond the reason people may have to care about preserving any other species). [↑](#footnote-ref-6)
7. This scenario may seem so improbable that it is not worth serious consideration, but in fact many scientists find it almost inevitable that such advanced artificial intelligence will be developed this century (Dietrich 2007) or that humans and machines will ultimately merge (Lovelock 2019). Bruce Tonn and Nick Bostrom also seem to take this possibility seriously writing, respectively, that “if future generations of [artificial life forms] deserve the same ethical consideration as future generations of humans, then their numbers need to be added to the population estimates” (Tonn 2009: 434) and “[s]uperintelligent machines might be built and their actions could determine the future of humanity” (Bostrom 2009). [↑](#footnote-ref-7)
8. Bostrom, again, is explicit in his priority for technological maturity and transhumanism over the existence of human beings. [↑](#footnote-ref-8)
9. Thank you to H. Orri Stefánsson, an anonymous reviewer, and audiences at the Global Priorities Institute at the University of Oxford, the German Association of Philosophy Colloquium at Humboldt University, and the University of Manitoba for helpful comments that improved the paper. Work on this paper was generously supported by Riksbankens Jubileumsfond and an Insight Development Grant from the Social Science and Humanities Research Council of Canada. [↑](#footnote-ref-9)