Offsetting the Harms of Extinction[1]

MICHAEL DA SILVA
University of Toronto

ABSTRACT

Many people assume that the extinction of humanity would be a bad thing. This article scrutinizes this apparent badness and demonstrates that on most plausible consequentialist frameworks, the extinction of humanity is not necessarily bad. The best accounts of the badness of the extinction of humanity focus on the loss of potential utility, but this loss can be offset if it is the result of sufficiently large gains by the present generation. Plausible means of calculating the goodness of outcomes accordingly suggest hastening extinction even in some circumstances where the alternative is a long period of human existence at a high level.

Keywords Ethics, consequentialism, existential risk, harms, extinction

INTRODUCTION

Many fear the potential extinction of humanity due to the common intuition that extinction is bad and should be avoided.[2] Yet what it means for extinction to be 'bad' is not obvious. This article scrutinizes the apparent badness of extinction. The most plausible candidate explanations for the badness of extinction do not rely on extinction itself being bad but on extinction pairing with other negative effects or forestalling other potential goods. Not all extinction scenarios have these implications. Extinction is not an impersonal bad and need not be personally bad even if we grant potential persons some moral personhood. Extinction is thus not necessarily bad. Even imminent extinction may be preferable to the continued existence of humanity for

1 Thank you to Derek Parfit and Jeff McMahan for comments on the earliest version of this article, which was drafted for their graduate seminar at Rutgers University. Thank you also to the other students in that course for thoughtful conversations on many issues and to the anonymous reviewers for feedback on more recent drafts.

2 As Larry Temkin notes, "anything…anyone…writes on this topic should be taken with a large grain of salt" (2008: 193). It is hard to know what the futures below would look like. This may affect intuitions about some cases and the theories used to explain them. 'Extinction' here refers to the extinction of humanity. The argument has implications for other extinctions.

very long periods of time on plausible means of calculating the value of outcomes if the extinction is brought about under the right circumstances. Once one recognizes that the badness of extinction is reducible to this lost potential utility, confidence in the intuition that imminent extinction is a bad thing that is to be avoided and/or delayed can be challenged on most plausible forms of outcome analysis that take potential utility into account. The lost potential utility of even a large number of future generations living lives that are worth living could be less than the amount of utility accrued by the current generation.[3] Extinction scenarios thus do not give one reason to choose between competing theories of outcome valuation.

The argument for these claims consists of six substantive parts. The first section assesses competing theories of the good and demonstrates that the badness of extinction is reducible to the lost potential utility of future generations that could exist but for the extinction (and any negative effects on existing persons). The second section briefly canvasses the best means of calculating the value of potential utility and outcomes including potential utility. I argue that intuitions that extinction is a bad thing to be avoided and/or delayed are undermined regardless of which mainstream position one takes. On Total-, Average- or Perfection-based analyses, the badness of extinction can be outweighed if it takes place as a consequence of an act that creates sufficiently good benefits for existing persons. The third and fourth sections demonstrate that this is true in cases where there is a choice between extinction and humanity continuing to experience lives worth living for a short period and cases where the alternative to extinction is humanity continuing to exist with very good lives for very long periods. The fifth section examines the significance of potential future flourishing generations in the analyses of the badness of outcomes. The final substantive section further defends the approach to extinction above by highlighting how it explains a separate intuition that the death of the last person is not the worst death in the history of humanity.

1. APPROACHES TO VALUING EXTINCTION

The claim that extinction is bad could mean several things. This section presents several alternatives and demonstrates weaknesses with many of them by way of defending the relative plausibility of a particular view.

The most common view on the value of extinction is probably something like:

a. Extinction is intrinsically bad.

John Broome helpfully explains the structure of this view (but does not

3 'Utility' here refers to whatever is valuable in life. Those who are queasy about hedon-focused Utilitarianism can substitute their own units of measurement.

defend it) (2012: 180-181). Contrary to (a), however, there are cases where many would not want to avoid (even near-term) extinction. All-else-being-equal, it is implausible to deny that one should choose extinction now over a million years of people living lives not worth living. Moreover, this position seems confused on terminological grounds. Non-existence has no intrinsic features/properties.

If the badness of extinction is not intrinsic, it is likely tied to its effects on the amount of utility that is realized in the world. This raises issues in moral mathematics that can be fruitfully explored in extinction cases. One type of value assessment appeals to purely person-affecting principles in which the goodness and badness of outcomes is determined by their effects on persons. The most common response to extinction may be best explained by what Derek Parfit calls the Narrow Person-Affecting Principle, according to which one of two outcomes cannot be worse if it would be worse for no one (1984: 393-395). Common aversive responses to extinction likely stem from imagined links between suffering and extinction. In several plausible scenarios, painful deaths act as a prelude to full extinction. Common responses to such cases support:

> **b.** Extinction is bad because the effects on (including harms to) existing persons are sufficiently great to render it bad.

If extinction were to take place as a result of a disaster that brought untold suffering with few benefits, (b) would be true. But determining whether extinction is necessarily (even comparatively) bad or necessarily includes bad-making features requires sifting out contingent facts. Extinction need not include such effects on existing persons. An impartial non-human observer interested in utility could lament the suffering in (b), but this would not entail lamenting the fact of extinction. If the early deaths are the price for a shorter period of an extreme well-being greater than the amount of well-being expected for any potential extra years of life, it is plausible that early death would not harm them. Imagine an extinction case where this is true for all existing persons such that no one currently alive is harmed by extinction. (b) is false in such circumstances and many others. Further, Strict Person-Affecting Views, which calibrate the goodness of outcomes using the effects on persons alone, tend to have unintuitive determinations about the supposed badness of extinction. On these views, extinction is bad iff the sum of utility of existing people lost by the act of extinction is larger than the sum of the utility gained. This does not always hold true. Therefore,

extinction is not always bad here.[4]

Issues with (b) lead theorists to seek other ways in which extinction is a bad thing. One attempt merely modifies the Person-Affecting Principle. This implausible approach can be dealt with briefly. The badness of extinction is often thought to go beyond its effect on the currently living. Some thus suggest that extinction is bad because of its effects on future persons. They explain the badness of extinction by extending the scope of the Person-Affecting Principle to include potential future persons who do not exist because humanity goes extinct prior to their birth. On such views, the badness of extinction can be calculated by some mixture of the effects on existent and potential persons, resulting in views like:

c. Extinction is bad because the effects on potential persons (which do not include direct harms) are sufficiently great to render it bad.

and

d. Extinction is bad because the effects on existent and potential persons are sufficiently great to render it bad.

These views likely describe the common views of lay persons, but are mistaken. At the time of extinction, potential future persons do not exist and cannot be harmed in the person-affecting sense.[5] If potential persons cannot be harmed, future generations are not directly harmed by extinction either. It is not, then, enough to appeal to person-affecting principles about what might be in or against the interests of presently existing people and future people when analyzing outcomes if one wants to salvage the intuition that extinction is always bad. The modified Person-Affecting View nonetheless hints at an important point: there is reason to take future generations into account when making moral decisions today and the sense in which future persons are morally relevant explains why we should usually avoid/delay extinction.

It is more plausible that the badness of non-existence stems from the fact that the history of the world would be better if extinction came later or never came about. The badness of extinction is impersonal. Jeff McMahan

4 Complex Person-Affecting calculi better demonstrate the potential badness of extinction. James Lenman (2002) suggests we care about future generations for selfish reasons (e.g., joy of knowing about future generations analogous to the joy of having children). This construction includes a personal bad, but hardly supports the idea of extinction as an impersonal bad whose badness extends beyond its effect on persons. Samuel Scheffler (2013) suggests that the badness of extinction partly stems from the way that it negates our ability to value and thus destroys utility in the present and future. These contingencies may be undermined in some cases below.

5 They will not exist unless we act in certain ways. Slight policy changes produce different future persons. Recall Derek Parfit (1984: Chapter 16).

(2013) plausibly ties together this impersonal bad and the potential interests of future persons. He suggests that the non-existence of a potential person is an impersonal loss. One cannot care for these persons morally for their own sake. McMahan nonetheless holds that one has a reason to bring a better off person into existence rather than a worse off person, which he suggests implies a reason to bring the better off person into existence rather than no person at all. To bring a person into existence is to confer a "non[-]comparative" benefit on him/her (9). Extinction is potentially problematic because it forestalls the granting of many non-comparative benefits and thus produces a history with less utility than a history in which extinction either never takes place or comes much later and non-comparative benefits are bestowed on new persons. The most important implication of McMahan's view for the extinction case is that there are impersonal reasons to bring people into existence due to the value they will add to the world. The perspective of the aforementioned impartial non-human observer interested in utility is the best point of view from which one can assess the potential badness of extinction. From this perspective, extinction is bad because it forestalls potential utility. Potential persons do not lose something by failing to come into existence. Instead, if causing people to exist would be good for them, their not coming into existence is bad despite not being bad for them.[6] If these people could have had lives worth living, their non-existence is an *impersonal* loss of value. The lack of benefits is a detriment in the history of the world. Comparisons of the utility of worlds with future generations and those without them help identify the bad of extinction: potential utility is not realized in the world where extinction is earlier.

One should, then, count the potential future utility of presently non-existent people when choosing between outcomes. This is not because of a duty to potential persons or because existence would be good for them. It is because it is comparatively better to have more utility in a given history than less utility. All-else-being-equal, it is better to bring about an outcome that realizes more of what is now merely potential utility than one that realizes less of it. If we count potential harms in our calculus of the badness of extinction, two plausible views arise. Given the contingency of an extinction scenario harming current individuals, one may adopt a view focused on impersonal loss alone:

> **e.** Extinction is *comparatively* bad if the loss of potential utility that would have accrued had the currently living people existed for a longer period of time and had other persons lived in the future is greater than 0.

6 Our "moral reason to ensure the existence of future generations is at least in part a moral reason to provide, or not to prevent, the enormous benefits of life for the enormous number of people who might exist in the indefinite future" (McMahan 1986: 335).

Yet even the truth of (e) depends on how extinction arises. Those interested in utility more broadly should take account of the utility of existing persons as well. While the badness of extinction may be reducible to (e), full utility-based outcome analyses cannot ignore when an outcome includes the suffering of current existing persons; contingent suffering is relevant when present. The impartial non-human observer cannot ignore it. An alternative thus combines person-affecting and impersonal perspectives:

> **f.** Extinction is comparatively bad if the sum of the suffering it imposes on living persons, the loss of potential utility that would have accrued had the currently living people existed for a longer period of time and had other persons lived in the future, or some combination of the two is greater than 0.

Richard Kraut, an opponent of absolute/intrinsic value, supports something like (f). For Kraut, the extinction of any species is bad iff the loss of the species is bad for the Earth's other creatures (2011: 169). The loss of beings that can and do experience and produce more good is worse than the loss of being who can and do experience and produce less good (185). Humans, including future humans, can experience and produce more good than any other species. Thus, the extinction of humanity would be the greatest of all catastrophes (164). Like McMahan, Kraut takes future generations into account when assessing outcomes. He thereby commits to a view whereby potential utility must be weighed in our moral calculations. Both the last generation of humanity and their possible beneficiaries in future generations would be negatively affected by an extinction scenario, reducing total utility in the world (164-165). Occasionally, Kraut makes it sound as if future generations could be harmed by extinction, but to the extent that he can be plausibly be read as endorsing (f), his view appears more plausible than alternatives.

Extinction scenarios, then, are most likely bad because of their negative impact on existing persons (to the extent that such effects are present) and because of the loss of the possible goodness of the people who might have existed and had good lives. The following explains how to compare the values of histories including the potential utility of future persons and how plausible calculations still lead to scenarios where extinction today is preferable than many years of continued human existence. It thereby explains why one should adopt a new approach to the badness of extinction, which is introduced in the next section.

2. CALCULATING THE BADNESS OF EXTINCTION

There is, then, a comparative harm in future people failing to come into existence if they would experience utility that would not otherwise be realized.

This harm is impersonal. If the badness of extinction is comparative and its value is exhausted by the loss of potential utility (and perhaps the disutility experienced by existing persons when the extinction scenario arises), this raises questions about how to calculate potential utility and the overall utility of an outcome.

The loss of potential utility stemming from an early extinction is a bad-making feature of an extinction scenario. Comments on *how bad* it would be are necessarily speculative,[7] but an impartial non-human observer would likely possess better measurement tools than I do. This piece thus assumes determinations on how much potential utility future persons would realize if brought into existence can be made, bracketing one source of uncertainty in population ethics, uncertainty about value, to assess the badness of extinction.

One cannot place potential utility valuation completely in a black box, but attempts to answer hard questions about such valuation raise several problems. Practical decisions rely on information available to modern humans, not impartial non-human observers. When comparing potential histories, we want to know if, for example, we should discount benefits to future persons or if potential utility is equivalent to actual utility (see e.g., Bostrom 2002: 15-16). Black boxing may thus be practically problematic. For present purposes, however, it suffices to note that the loss of potential utility is non-negligibly bad.[8] Regardless of how one values potential future

7 John Broome agrees that we must take potential persons into account (2012: 175). The absence of persons accounts for our intuitions about the badness of extinction, even if we do not think it can explain why we think extinction is any worse than any other massive drop in the potential population. Yet Broome is more skeptical than McMahan about the ultimate badness of large absences of persons:

> Intuitively it seems most plausible that…[absences] are bad….But…we still have a lot of work to do before we can be sure that this is so….[E]ven if we can be sure a collapse of population would be bad, we have no idea how bad it would be. We have empirical work to do in predicting what would have been the well-being of the absent people, had they lived (183).

This provides reason to question (e), (f), and (g). Broome suggests we cannot be sure of our utility calculations and thus may not be able to do the moral mathematics necessary to support the views. If this is true, any comments on the value of potential utility, including comments on the value of outcomes that rely on potential utility calculi, are necessarily speculative.

8 I am tentatively wont to provide such a discount based solely on the uncertainty identified by Broome, but the claim that the badness of extinction can be outweighed by other relevant circumstances even when the alternative would be many years of continued human existence does not depend on such a discount. E.g., Parfit offers a Two Tier View, according to which we give greater weight to the badness of outcomes that would be worse for particular people, but give some weight to non-person-affecting good and bad outcomes (2011: 219-233). Questions concerning whether extinction is always bad and whether we should always attempt to delay it arise even on versions of the Two-Tier View that give significantly less weight to effects on future well-being that do not affect particular people.

utility, it should be included in assessments of the good of outcomes. The more pressing concern is how to calculate the overall utility of an outcome given fixed inputs of the utility of existing persons and potential utility of future persons.[9] Two popular candidate principles for such determinations are the Total Principle and the Average Principle. The former holds that "other things being equal, the best outcome is the one in which there would be the greatest quantity of whatever makes life worth living" (Parfit 1984: 387), but unfortunately entails the Repugnant Conclusion (388). The latter holds that the best scenario is one in which the average amount of utility experienced by each person is highest and may have similar implications in its widest form (399). It is also subject to further critiques, including the Levelling Down Objection (described in Temkin 2012: 75-76). There is thus reason to question the most intuitive Non-Person-Affecting Views. Nearby views suffer from similar defects[10] and extinction cases like the ones below raise similar problems.[11] All principles of valuation suffer from some defects and are thus not obvious candidates for use in the valuation of the badness of extinction.

The following possibility, which is agnostic about the competing principles, helps avoid these problems, but also supports the view that extinction is not necessarily bad:

> **g.** Extinction is not comparatively bad if the sum of any negative disutility experienced in the process bringing about extinction and the impersonal negative effects of the potential utility of existent and future persons failing to be realized can be negated by earlier benefits conferred on existing persons.

The scenarios below suggest followers of Total-, Average- and Perfection-based outcome valuation principles should all prefer imminent extinction

9 For simplicity's sake, calculations here ignore Different People Choices, wherein different persons will be born depending on which of two scenarios arise and we assess the relative value of their lives (Parfit 1984: 356). The choice is between only this generation existing and any future generation existing.

10 Given space limitations, other principles cannot be canvassed. Yet it should be reasonably clear that nearby view suffer from similar defects. E.g., those who understand the case demonstrating how the Average Principle may lead to Repugnant Conclusion should understand how these arguments also apply to the Average Utility Principle. Small differences in particular cases are dealt with briefly below. The key is that the treatment of (e)-(g) above remains true when reformulated to account for average utility.

11 The Impersonal Total and Average Principles also entail that, under certain circumstances, extinction is preferable to long periods of continued human success. Indeed, the cases below suggest that extinction may be preferable on any plausible valuations. Given the similar problems between these views and their nearby alternatives on the margins, it is likely that the total badness in (e) and (f) can be negated by earlier benefits conferred on existing persons regardless of whether the loss of utility is calculated in totals or as deviations from an average.

provided that the limit on the amount humans are able to flourish is sufficiently high. (g) is thus true regardless of whether one calculates the value of outcomes from a Total-, Average- or even Perfection-based perspective.[12] Given that the most plausible outcome valuation theories all rely on one of these principles, one should not choose a theory solely to account for one's pre-theoretical intuitions that extinction is bad and should be avoided or delayed to the greatest extent possible. Regardless of whether one assesses the comparative badness relative to the possible total sum of utility that would have been contained in the lives of people who would have otherwise existed, on their quality of life, or some combination of these, imminent extinction may be preferable to long continued periods of human existence even at a high level of well-being.

This does not entail that extinction is always the better outcome, but only that an early extinction may be a better outcome than a later one (from an outcome perspective) and a history with extinction in it may be preferable to one without it. This is an argument against those who consider extinction to be intrinsically bad and argue that it is always the worst, including those who say it would be intrinsically worse than humanity's continuing to exist for longer.[13] The main arguments for this claim are case-based and appear below. Following theories to their logical extremes to derive implausible results is common in ethics. I hope to show that any view on valuation may have the implausible result that extinction could be preferable to continued human instance. This is not meant to be an argument against consequentialism, but it should help demonstrate that one should not accept a particular form of consequentialism just to avoid the conclusion that extinction is preferable to alternatives.[14] For instance, McMahan uses the badness of extinction as a datum for why one should admit non-comparative benefits, the aforementioned benefits that "cannot be explained in counterfactual comparative terms" (2013: 9), into one's moral mathematics (26). For McMahan, extinction appears to be "the worst of those possible tragedies that have more than a negligible

12 Perfectionists believe ensuring people have a high quality of life is most important. Perfectionism too produces results where extinction is preferable to even long periods of continued human existence. E.g., the impartial perfectionist who is only concerned with the potential humanity being fully realized may prefer a world in which humanity flourishes to the greatest extent possible now even if the non-existence of many future generations who would otherwise exist is a necessary consequence.

13 I will not address an extreme view one could read into David Benatar (2006: 194): h. Earlier extinction is preferable to a later one because coming into existence is always harmful. We are obligated not to harm people and thereby obligated to hasten extinction by not procreating.

14 If (g) is true, those who believe that extinction is necessarily bad need to look outside utility-based analyses for justification. A rule-based approach to ethics may justify this belief. Adopting such an ethics may be the right application of the argument's conclusion. This piece merely seeks to identify implications of utility-based analysis.

probability of actually occurring", not merely due to its effects on existing persons, but also due to the loss of potential future utility by potential future persons (26). Since potential future persons when choices concerning extinction are being made, there may be no relevant counterfactual in which they are comparatively benefitted or harmed. The purported losses of extinction thus appear non-comparative. Extinction produces impersonal losses. McMahan's view's ability to explain the general plausibility of (e) and (f), in which extinction will almost always be at least comparatively bad, counts in its favor. The loss of potential value in (e) and (f) are best understood as non-comparative or impersonal. One should not, however, assume that (e) and (f) are true. Indeed, even McMahan's mathematics can be used to create a choice scenario where extinction is not the worst outcome. Plugging non-comparative harms into (e) and (f) can still result in 'Extinction is bad' reading false. The badness of extinction alone thus does not justify admitting non-comparative benefits and harms into our moral calculations. McMahan is aware of other problems with non-comparative benefits and harms, but these considerations suggest that the extinction case may not provide adequate reason to accept them in the first place. To the extent that one prefers one's intuitions about the badness of extinction to one's ability to make plausible moral calculations, this is a problem with utility-based theory. Others should be moved to reconsider their distaste for certain imminent extinction scenarios. The remainder of this piece will demonstrate that one should not admit the potential utility of future persons into one's moral calculations merely to explain pre-theoretical intuitions about the badness of extinction. This is because the addition of these people into our moral calculus will not always allow us to maintain these intuitions. Providing future individuals with the means to realize their potential utility is good. Since this good is merely comparative, however, it is not morally necessary that one bring it about in all cases. Since it is impersonal, no one is harmed by failing to realize it. When the potential utility calculus is paired with the most plausible means for analyzing the overall goodness of outcomes, the loss of potential utility of a hastened extinction will not rule out choosing extinction over histories where human beings live longer in certain circumstances. Extinction today may be preferable to millions of years of continued human existence in some circumstances.

3. EASIER CASES

The extinction of humanity, then, is not intrinsically bad and any potential negative effects on existent and potential persons can in principle be negated by earlier benefits conferred on existing persons. The following cases, focused on the use of pills that are unavailable in the physical world but

common in philosophy, support the more fundamental first conjunct concerning the intrinsic value of extinction. One may prefer a history with an earlier extinction to a latter one and a history with extinction in it to an alternative without it. Moral mathematics does not always demand choosing an outcome that avoids extinction. Nick Bostrom notes that it is not "a *conceptual* truth that existential catastrophes are bad or that reducing existential risk is right" (2013: 24). If one is solely concerned with outcomes, it also may not be a substantive truth that extinction is necessarily bad. Harder cases below suggest early extinction may be preferable to circumstances in which humanity survives for a very long time. I first address less contentious cases where humanity will only continue to exist somewhat longer.

The supposed badness of extinction is often demonstrated with hypothetical scenarios, but such scenarios also undermine this supposed badness. Larry Temkin provides an example of a scenario in which mass sterility leads to extinction to suggest that an outcome where regular regeneration continues is better than one giving current people immortality; contra Jan Narveson,

> if we developed a pill enabling each of us to live wonderful lives for 120 years, it would be terrible for us to take the pill if the cost of doing so were the extinction of humanity. This is so even *if* taking the pill were better for each individual who took it, and hence everyone whoever lived, collectively.…[I]f the cost of immortality would be a world without infants and children, without regeneration and rejuvenation, it wouldn't be worth it.…[T]his is so even *if* each immortal would be better off than each mortal (2008: 208)[15].

Intuitions about similar cases are supposed to demonstrate the badness of extinction. Yet I suspect that our intuitions about the case will differ if it is altered such that existing persons are made sufficiently well off. Extinction may be the worst outcome of a given decision, but if we remove personal harms from the scenario, extinction can be personally good. In such circumstances, the impersonal loss is merely a function of the lost potential utility of future generations that would have otherwise existed. A sufficient level of personal good for existing persons could outweigh this loss.

From a pure outcome perspective, case-based reasoning suggests that a history including extinction may be preferable to an indefinitely long history without one.[16] Imagine a choice between:

15 My thoughts on this topic were furthered by two Temkin-inspired cases in Nick Beckstead's doctoral dissertation (2013: 63). Gregory Kavka provides another famous pill case (1982: 98).

16 These intuitions affirm Lenman's claim that "[f]rom an impersonal, timeless perspective it is hard to identify good reasons why it should matter that human extinction comes later rather than sooner" (2002: 253).

The Highest High: An intergalactic travelling salesman arrives on Earth. The salesman offers the Earth's inhabitants a pill that allows everyone currently alive to reach the highest level of flourishing possible. Infertility is a side effect. The salesman is only on Earth for a brief period of time and will not make the offer again, but will only provide it to the current generation on the condition that everyone agrees to take it. Everyone agrees to take the pill. Humanity goes extinct when the last currently alive person dies.[17]

Rejecting the Offer 1: The intergalactic travelling salesman makes his offer, but it is rejected. Humanity continues to develop, but extinction comes within a few hundred years due to natural circumstances.

Rejecting the Offer 2: The intergalactic travelling salesman makes his offer, but it is rejected. Humanity development plateaus due to unforeseen technological problems. Extinction comes within a few hundred years due to the natural circumstances from Rejecting the Offer 1.

Return to the Repugnant Conclusion: The intergalactic travelling salesman makes his offer, but it is rejected. Human development regresses. A large number of humans continue to exist for an indefinite period of time with lives barely worth living.

The pill's extraterrestrial origin removes contingencies in other pill cases.[18] Many of the worries surrounding extinction are also removed. Preferences can be satisfied. Voluntariness is not undermined. Even the violent ends of the last generation that add to the badness of extinction in similar scenarios are not present.[19] Most forms of uncertainty are removed from the comparative equation. The possible outcomes are stipulated to identify whether one with the extinction of humanity in it is necessarily

17 For simplicity's sake, assume that the last people die together, everyone enjoys full material comforts, and no family members see each other suffer. This avoids pains in Lenman (2002: 255).

18 In the absence of an 'all or nothing' decision on whether to take the pill, it is best to delay taking it until either scientists develop it without the sterility side effect or it is clear that the side effect could not be remedied. It remains important to determine whether extinction following flourishing is problematic rather than focusing on when one can know the following periods will not be better. If the side effect could not be remedied, the case would be akin to the extraterrestrial introduction in all relevant respects.

19 Lenman provides a famous example of such a scenario and poses two questions: Suppose it is written in The Book of Fate that one day we will be wiped out in a nasty catastrophe. Many millions of people will die in terrifying circumstances involving great pain and distress. The only thing the Book of Fate is silent about is when this is going to happen….The question is – Should we care? Does it matter how soon this happens? (2002: 255).

worse than the alternatives.[20] It is not obvious that The Highest High is the worst scenario. It is thus not obvious that extinction sooner rather than later is necessarily a bad outcome. Human beings' ability to flourish could be limited by their nature and psychology. If so, a relatively small number of future generations existing below the limit may produce a larger number of positive benefits than the pill. If, however, the level of flourishing is sufficiently high, then The Highest High creates more utility than Rejecting the Offer 1 and 2. It thus appears to be the preferable outcome.

The choice above may be a mere choice between existential risks,[21] but this does not undermine the broader implications of the example. When compared with Return to the Repugnant Conclusion, the mere presence of extinction in the Rejecting the Offer scenarios does not make the situation worse than an alternative without it in any substantial way.[22] Nick Beckstead (2013) is likely right that a given period with people in it is better than a period without sentient life, but the preceding choice scenario suggests that the disvalue of empty periods can be outweighed by sufficiently good periods when we look only at histories.

4. HARDER CASES

One may charge that the important comparison involves not just a few more centuries, but a much longer survival of humanity. Parts of Parfit's *Reasons and Persons* (1984) and other influential works in population ethics assume that the human race could continue to exist for a long time.[23] They then question whether an earlier extinction would be preferable to such long histories. Even those who prefer the Highest High to Rejecting the Offers 1 and 2 would likely find it less obviously preferable to a future where human beings continue to live for longer periods.

20 Broome's uncertainty about value potentially remains. This lingering uncertainty about the extent to which things are good or bad is no worse than what we find in any other scenario. Even Broome notes that expected value theory will not help with this uncertainty (2012: 184).

21 E.g., Rejecting the Offer 2 includes a long period of stagnation (which is not Nick Bostrom's "permanent stagnation" (2013: 20) since extinction occurs).

22 One may argue that this would result in a decrease in morally relevant diversity, but a fully realized human contributes to diversity in the history, resulting in a tradeoff of the loss of diversity. There is reason to question the long-term relevance of this diversity criterion even in the absence of that tradeoff. As Lenman argues (2002: 255), it seems more important that humanity exist at some point in a history to contribute to diversity than for it to continue to exist indefinitely. Diversity could be a bad-making feature of extinction at any given time, but if we take a sufficiently impartial view and analyze outcomes of whole histories, it is no longer relevant. Diversity over a history may additionally benefit from humans failing to exist if some species can only exist where humans do not.

23 Bostrom suggests this is an issue with many existential risks (2013: 22).

It is important to examine these harder cases comparing early extinction to a history where humans continue to live for longer periods. Yet the only fundamental difference where one is using the most plausible outcomes valuations is that the amount of utility the current generation would need to experience to make imminent extinction preferable is much higher than it is in the easy cases. Even if we grant that the loss of potential persons could make a history worse, extinction is not worse than even alternatives where humans continue to live even very good lives for thousands or even millions of years if it came about as a consequence of existing people being guaranteed lives that were very much better.

Consider:

> Rejecting the Offer 3: The scenario in Rejecting the Offer 1 takes place but thousands of years pass before the extinction of the human race due to natural circumstances.
>
> Rejecting the Offer 4: The scenario in Rejecting the Offer 2 takes place but thousands of years pass before the extinction of the human race due to natural circumstances.

Given a sufficiently long period of time, one may plausibly believe that the gains in quantity of lives in these outcomes when compared with the Highest High would be outweighed by the lower quality of people's lives. Much longer time periods make hastened extinction less compelling.[24] If one accepts Beckstead's claim that "it is not absurd to consider the possibility that civilization continues for a billion years, until the Earth becomes uninhabitable" (43), Rejecting the Offer 3, in which humans continue to develop, or Rejecting the Offer 4, where human development plateaus, could be plausible constructions of these long histories.[25]

Variations on Rejecting the Offers 3 and 4 suggest that extinction should often be avoided, but, given certain assumptions, the Highest High may still be preferable. From an outcome-based perspective, extinction should be avoided where the positive benefits of an act that will result in or hasten extinction will not outweigh the loss of potential utility of future generations. For any given comparison with a potential future, one should focus on the potential utilities of future histories. To determine whether the Highest High is preferable to Rejecting the Offers 3 and 4, one must be able to calculate the total amount of utility in each. Whether the Highest High will outweigh

24 One may worry that these additional numbers will eventually lead to Return to the Repugnant Conclusion. The structure of Rejecting the Offers 3 and 4 ensures a relatively high amount of well-being in both scenarios. I nonetheless discuss this concern below.

25 Return to the Repugnant Conclusion is unlikely. Broome says "we cannot reduce the chance of extinction to zero" (2012: 179). Richard Kraut agrees (2011: 163). They are likely right.

Rejecting the Offers 3 and 4 depends in part on what 'the highest level of flourishing possible' in the Highest High means. It is easy to see how, all-else-being-equal, much longer periods of time will create much greater amount of utility over the history of humanity. Rejecting the Offers 3 and 4 thus include more utility than Rejecting the Offers 1 and 2 respectively. Whether they will include more utility than the Highest High is not obvious. It is natural to assume that we will eventually reach a point where the amount of time is sufficient long that even a much smaller amount will sum (or even average) to a higher amount than the pill in the Highest High could possibly reach. Return to the Repugnant Conclusion is supposed to make this clear. In cases where extinction will eventually take place, albeit millions of years later, the question of whether more people experiencing less good for longer periods of time includes more utility than everyone alive today experiencing the highest amount of utility possible depends on how much utility the present generation could enjoy. It is hard to imagine 'indefinite utility' that could offset any potential lesser good in the future. There likely is a limit to the amount of utility any person could experience, but (g) remains true where the alternative history extends for thousands or even millions of years iff the limit on the amount of utility currently existing persons could accrue is sufficiently high that they could accrue more utility than many future generations. If the limit is sufficiently high, it may be such that the good current persons get from taking the pill is greater than thousands, millions or even billions of years of existence in any of the four Rejecting the Offer scenarios.

If the gap between present utility levels and our maximal utility levels is sufficiently high and one is only interested in choosing between better outcomes, then, one may choose the Highest High over Rejecting the Offers 3 and 4. Given what we know about human physiology and psychology, the gap between humanity's current utility level and the maximum amount we could enjoy is likely insufficiently large to offset millions of years at current or even lesser levels of utility. But imagine a pill that brings us beyond our current maximal capacity such that the highest level of utility is beyond current human limitations and results in each of us experiencing bliss much greater than the cumulative well-being of hundreds of persons at our current level living long lives. If this is the pill on offer in the Highest High, humanity would not err in collectively agreeing to take it on risk of sterility. Even if humanity would continue to develop such that future generations would flourish much more than we do today, experiencing goods far beyond our current capacities, a pill that could take existing persons beyond that level and provide the maximal amount of utility possible could produce more utility provided that the maximal amount of utility possible is sufficiently high.

One may suggest that beings who took that pill would no longer be recognizably human. The pill would then result in the immediate extinction of humanity by another name. Yet most theorists agree that any history of humanity that will continue for thousands, let alone millions, of years needs to appeal to human beings' descendants (e.g., Beckstead 2013: 43). The relevant comparison thus assumes we are dealing with beings that may not be recognizably human (but are closely related). While some will reject this assumption, it is sufficiently pervasive to support my demonstration that there is a way of understanding the Highest High that makes it preferable to Rejecting the Offers 3 and 4.[26]

If the limit of human flourishing is sufficiently low, the Highest High may not be preferable to different Rejecting the Offer scenarios. Eventually there will be a long enough period of time that will make rejecting the pill necessary given a sufficiently long period of time and some cap on the highest level of utility possible. A problem for this salvation of anti-extinction-based intuitions nonetheless threatens. Perhaps any time extension of this sort would create a gap between the level of well-being of pill takers and future generations such that the scenario would mirror Return to the Repugnant Conclusion in certain respects. The idea that a world with more persons who are less well-off could be better than a world that has a smaller but still considerably large number of persons (and, indeed, more than enough for society to function) who are much better off strikes many as implausible, but the source of the repugnance of the repugnant conclusion is hotly debated. The Repugnant Conclusion seems to demonstrate that, once a sufficient number of persons exist, the aim should not merely be to ensure more people exist, but also to ensure that each person experiences a certain level of well-being. Given that the persons in both worlds are living lives worth living, the problem cannot be that the level of well-being in either world is too low in an absolute sense. The repugnance of the Repugnant Conclusion only occurs in comparative analyses. One explanation for it is that the gap in well-being between persons in the first possible world and those in the other is too large to be justifiable. The gap in quality of life across worlds makes the creation of lives worth living seem repugnant even when the lives would otherwise be worth living. If Return to the Repugnant Conclusion is problematic not because of the much lower amount of well-being allotted

26 Depending on how one individuates species, it is possible that Beckstead's multiple phases of humanity/post-humanity will contribute to diversity more than the instant development of the pill in the Highest High. One may argue that this would be a further bad-making feature of extinction in this circumstance that is not adequately covered by utility calculations. Even if one grants that the manner of species individuation that would undermine my position is correct, it is possible that the number of species that could flourish in the absence of humanity would be greater than the number of post-human species. Such speculation should be examined elsewhere.

to each person in the world with more persons as such, but because it is much lower than what we take to be acceptable, perhaps a sufficiently high level of maximal utility could make existence even at the a very good level seem repugnant compared to the blissful level produced by taking the pill. Even the existence of many more persons for a long period of time at a current level of well-being may seem repugnant when one compares the quality of life at the blissful level with the quality of life at our current levels of well-being. Rejecting the Offer 4, where human development plateaus, seems particularly problematic here, though development at a slow enough pace in Rejecting the Offer 3 could also be worrisome. I suspect that the comparative explanation for the repugnance of the Repugnant Conclusion is the wrong tack, but the fact that the lives in Return to the Repugnant Conclusion are worth living makes the claim that they are absolutely, rather than relatively, bad implausible. Defenders of Rejecting the Offers 3 and/or 4 need to explain why we prefer the Highest High to Return to the Repugnant Conclusion without appealing to the large gap in the relative well-being of persons across the scenarios or risk a similar gap in the relative well-being of persons in the Highest High and Rejecting the Offers 3 and 4 undermining their position. The badness of extinction is still not as obvious as it seemed.

If we can limit the maximal amount of utility that could be brought about by the pill, lengthen the amount of time in the Rejecting the Offer scenarios to a sufficiently long period that the total utility in the scenario would be greater than that amount, and explain why Return to the Repugnant Conclusion is worse than the Highest High without appealing to a comparison that is mirrored by any Rejecting the Offer scenario and the Highest High, then it is easy to construct scenarios where even one who is only concerned with total utility in an outcome should refuse to take the pill. The number of conditions here would, however, likely surprise many. Laypersons likely believe their intuitions that hastening extinction is a bad thing will survive most scenarios. This jolt to intuitions strengthens the claim that the Highest High reveals a non-obvious truth about the badness of extinction on outcome-based analyses: it is comparative and can be offset.

5. POSSIBLE FUTURE GENERATIONS WHO WOULD GREATLY FLOURISH

If Beckstead's speculation about the future is correct, however, it is more likely that anyone who will approach the blissful level will do so through a gradual process of development (like in Rejecting the Offer 3). The intergalactic salesman is unlikely to arrive soon. Even if s/he could exist, it is likely that s/he will only visit in a far future in which we can communicate with extraterrestrials and interstellar commerce can be done efficiently. It is more likely that the highest level of flourishing will require continued technological

development.[27] It is, in other words, unlikely that we will flourish more than any potential future generation that could exist.

It is thus worth considering what we should believe about the possible existence of people whose lives would be vastly better than the lives of the most fortunate actual people, but a few comments will have to suffice here. If an impartial observer knew that the Highest High would take place 1,000,000 years from now, then, all-else-being-equal (e.g., assuming there are no periods where everyone has lives that are not worth living in the interim), s/he would have reason to prefer a history that lasted that long.

Consider:

> Weak Batch: The pill from the Highest High is offered to humanity in a diluted form that will only bring the existent generation up to level of the best life anyone is currently living. The salesman says s/he could provide a better batch in the future that would bring a future generation up to the Highest High. Ingesting the weak batch now will produce infertility that would make such a trip useless. Humanity takes the weak batch.

The value calculations above suggest humanity should not take a pill that could raise all existing persons up to the level of the best currently existing persons with the same infertility side effect as the pill in the Highest High if it knows that a much higher level of flourishing could be experienced by a future generation. The future generation would not be harmed by not being able to take the pill, but the history of the world would be worse if they were given the opportunity. Even a massive boost in well-being for the current generation beyond what anyone experiences today cannot justify hastening extinction to an earlier date. A 'Stronger Batch' situation produces the same result. From an impartial perspective, the current generation has no special standing.

Yet more interesting questions arise when we contemplate future periods of great levels of flourishing below the maximal level in the Highest High. Consider:

> Good Times Ahead: Development in Rejecting the Offer 3 creates a period of overwhelming positive utility in the future, much higher than

27 Bostrom posits a Technological Completion Conjecture: "If scientific and technological development efforts do not effectively cease, then all important basic capabilities that could be obtained through some possible technology will be obtained" (2009: 190). One can imagine a version that brings humanity to its highest level of flourishing. Given certain technological developments, we may reach a point where humanity's highest ends can be realized even without the intervention of an intergalactic traveler. If reaching this point requires sterility and we know this side effect is unavoidable, the choice scenario is similar to that of the intergalactic traveler. This piece provides guidance on how to make that choice. See note 18.

the level any human experiences today.

The forgoing provides the tools necessary to decide whether one should prefer this to the Highest High. If its "overwhelming utility" is greater than one would get by taking the pill, Good Times Ahead is preferable to the Highest High. If the maximal level of utility in the Highest High is sufficiently high and the "overwhelming positive utility" in Good Times Ahead is less than the maximal level, it is possible that the gap is such that even the addition of other periods could not result in utility at the level of the Highest High. The Highest High would thus be preferable.

6. EXPLAINING ANOTHER INTUITION

Extinction, then, is not bad in certain circumstances on most plausible outcome-based analyses. This helps explain common intuitions about the relative badness of the deaths of the last person and others. Many do not think the death of the last person is worse than the death of others who preceded him/her. The fact that one death would bring about the extinction of humanity is not seen as conclusive proof that it is worse than others. The simplest explanation for this intuition that does not run afoul of other plausible ethical stances is that the outcome of this death, extinction, is not worse than the outcome of other deaths where other persons remain.

The intuition about the relative badness of deaths is most easily raised when comparing the death of the last human and the death of the human immediately preceding him/her. It is stronger where we imagine that the last humans know each other. Many people do not believe that the last human death would be worse than the penultimate human death. The penultimate death may even be worse since the last person will mourn the penultimate person's death in the circumstances, if s/he knew that person, and then live alone without interpersonal connections that provide most of life's meaning. S/he could be deeply affected by the death of the penultimate person even if s/he did not know the penultimate person, but only knew of his/her existence. Samuel Scheffler "would choose not to live on as the only human being on earth even if the alternative were not that human society would survive after my death but rather that everyone including me would die… [This preference most importantly] reflects the strongly social character of human valuing" (2013: 80). This claim is supposed to be evidence for the badness of extinction, but can support the claim that the death of the last human may not be the worst one. For Scheffler, knowledge of imminent extinction renders one's life plans meaningless and one's projects valueless. One's current values are likewise tied to the existence of other persons at the same time. We need other people to value our lives. If Scheffler is right, the death of the last person is less bad than the death of second last person.

Desires to ensure that one last person remains alive, even indefinitely, are thus curious.

The fact that the last and penultimate persons are among the last members of society obscures a larger truth: we often think that the fact that the last person alive is the last person alive does not make his death any worse than the death of many, and perhaps even any, other persons. The mere fact that s/he was the last person in existence does not make his/her life any more valuable than another. Barring circumstances in which the person's status as the last person was the result of virtues fully in his/her command, we often think that this status is arbitrary and could easily be otherwise. If this is the case, there is little reason to mourn his/her death any more than we would mourn the death of an equally valuable contributor to society today. The claim that the deaths are not worse than one another is slightly different from the claim that there is no reason to mourn one more than the other. The latter claim is trivially true if we consider mourning to be a strictly *post hoc* determination: there is by definition no one to mourn the last person on Earth after his/her death. We must instead examine the former question in an *ex ante* manner and compare which of two deaths we would prefer not to take place in certain circumstances. This determination is similar to one on which death is worse all-things-considered from the standpoint of the impartial observer judging outcomes. Many think neither death is worse than the other. Some believe that the death of the last person on Earth is better than the death of earlier persons in certain circumstances.

The easiest explanation for these intuitions, treating one death as worse than the other seems arbitrary, is not the best explanation. The relative badness of the deaths of two persons who are otherwise the same should not be determined by the order of their death. Reversing the order seems morally irrelevant. Intuitions about the relative badness of the deaths of the last and penultimate persons thus cannot be fully explained by the irrelevance of the moral order of actions. The order of actions affects their independent moral status elsewhere.[28] This could be true where the order otherwise seems to be an arbitrary distinguishing mark between two cases. The best explanation for intuitions supporting the view that the death of the last human on Earth is sometimes no worse than and even preferable to earlier human deaths is simply that sometimes the later death is preferable *despite bringing about the extinction of a species*. In other words, the best explanation is that the ultimate outcome of extinction is preferable to an alternative in which persons continue to live in limited circumstances. (g) helps explain intuitions about the relative badness of

28 McMahan plausibly argues that "the order does make a difference" in determining the permissibility of certain actions in the domain of abortion and prenatal injury (2006: 649).

deaths without appealing to questionable claims about the moral irrelevance of the order of actions. This provides further reason to accept it.

CONCLUSION

The extinction of humanity, then, is not intrinsically bad and might be comparatively bad only by being an absence of what would have been good. This absence can be outweighed by current goods. Thus, the extinction of humanity is not always worse than alternative possible futures. Even the imminent extinction of humanity may be preferable to the continued existence of humanity for long periods of time at high levels of well-being on most plausible valuations of outcomes provided that extinction takes a certain form. Methodologically, then, one should not choose a means of valuing outcomes merely to avoid imminent extinction. Extinction may be preferable in certain circumstances regardless of what view ones takes. The insights here, then, have methodological value. They should also help clarify *why* extinction should not be hastened now and when it may not be the worst outcome.

BIBLIOGRAPHY

Beckstead, N., 2013: *On the Overwhelming Importance of Shaping the Far Future*, PhD Thesis, Department of Philosophy, Rutgers University.

Benatar, D., 2006: *Better Never to Have Been*, Oxford: Clarendon Press.

Bostrom, N., 2002: "Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards", *Journal of Evolution and Technology* 9.

— 2009: "The Future of Humanity", in *New Waves in Philosophy of Technology*, ed. Jan-Kyrre Berg Olsen, Evan Selinger and Soren Riis, 186-216, New York: Palgrave McMillan.

— 2013: "Existential Risk Prevention as Global Priority", *Global Policy* 4: 15-31.

Broome, J., 2012: *Climate Matters*, New York: W.W. Norton & Company.

Kavka, G., 1982: "The Paradox of Future Individuals", *Philosophy and Public Affairs* 11: 93-112.

Kraut, R., 2011: *Against Absolute Goodness*, New York: Oxford University Press.

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

McMahan, J., 1986: "Nuclear Deterrence and Future Generations", in *Nuclear Weapons and the Future of Humanity*, ed. Avner Cohen and Steven Lee, 319-339, Totowa, NJ: Rowman & Allanheld.

— 2006: "Paradoxes of Abortion and Prenatal Injury", *Ethics* 116: 625-655.

— 2013: "Causing People to Exist and Saving People's Lives", *The Journal of Ethics* 17: 5-35.

Parfit, D.,1984: *Reasons and Persons*, Oxford: Clarendon Press.

— 2011: *On What Matters: Volume 2*, Oxford: Oxford University Press.

— 2013: "Death and the Afterlife", in *Death & the Afterlife*, ed. Niko Kolodny, 15-110, Oxford: Oxford University Press.

Temkin, L., 2008: "Is Living Longer Living Better?", *Journal of Applied Philosophy* 25: 193-210.
— 2012a. *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, New York: Oxford University Press.