DOI: 10.1111/phpr.13066

Philosophy and Phenomenological Research

ORIGINAL ARTICLE

Why prevent human extinction?

James Fanciullo[1,2]

[1]Lingnan University, Hong Kong, Hong Kong

[2]Hong Kong Catastrophic Risk Centre, Hong Kong, Hong Kong

Correspondence
James Fanciullo, Lingnan University, Hong Kong, Hong Kong.
Email: jmsfanciullo@gmail.com

Abstract

Many of us think human extinction would be a very bad thing, and that we have moral reasons to prevent it. But there is disagreement over what would make extinction so bad, and thus over what grounds these moral reasons. Recently, several theorists have argued that our reasons to prevent extinction stem not just from the value of the welfare of future lives, but also from certain *additional values* relating to the existence of humanity itself (for example, humanity's "final" value, or the value of humanity in itself). In this paper, I argue against these "additional value" views. Despite their initial appeal, these views will inevitably face conflicts between the additional values to which they appeal, and the value of the welfare of future lives. And, I argue, the views cannot plausibly resolve these conflicts. In contrast, these conflicts do not arise for a rival view, on which our reasons to prevent extinction stem just from the value of the welfare of future lives. I conclude that this gives us reason to prefer the latter view, despite the greater initial plausibility of additional value views.

It is possible that human beings will someday become extinct. Indeed, it is possible that beings in any way relevantly close to human beings—perhaps like "posthumans," or beings resulting from modifying and enhancing our biological makeup—will at some point have no possible future

existence either.[1] Many of us, and dare I say most, think that these possible scenarios would be extremely bad. But why we think it would be so bad seems to require some exploration.

Some of us think the badness would consist, at least in part, in the fact that humanity's ending would rob the world of a great deal of goodness. Not only would the total amount of goodness decline alongside the dwindling number of those existing in the final generation, we might think, the potential for so much *future* goodness would be destroyed as well. After all, if humanity had ended in, say, the year 1900, all of the goodness deriving from those who have existed since then would have been robbed from the world. It is difficult to even imagine the analogous point concerning the goodness derived from those who will ever exist in the future—or would have, if not for our extinction at some point along the line. Needless to say, this amount of goodness seems immense. So, at least on this line of thought, at least part of what would make humanity's extinction so bad is that it would rob the world of a great deal of present and *future* goodness. The immense value of humanity's future existence then gives us strong reason to think human extinction would be extremely bad, and to protect against this possibility as best we can.

Few of us, I think, would reject this line of thought. Our views become less uniform, however, when we ask about what *makes* humanity's future existence good. Obviously, some of us will claim, a major factor in what makes the existence of future lives good is that the lives are themselves good—that they are high in well-being or welfare, or at least that they are lives worth living. Taken to the limit, this line of thinking leads to the conclusion that the value of humanity's future existence *just is* the total amount of goodness contained in future lives. That is, what makes humanity's future existence good just is that the future lives would be good, and the goodness of this future existence is just a function of the number and quality of these lives (or, a function of how many lives there would be, and how good those lives would be). On this view, then, there is no further value contributing to the goodness of humanity's future existence, beyond the aggregation of the goodness of future lives. Call this *the welfare aggregation view*.

Those who reject the welfare aggregation view need not deny that the welfare of the future lives contributes anything to the value of humanity's future existence. They might claim that the welfare of the future lives is at least *part* of what makes humanity's future existence valuable. And indeed—or so it seems to me—they had certainly better: any plausible stance here will at least accept *that*. After all, while there may plausibly be other forms of value contributing to the goodness of humanity's future existence, the value of the future lives themselves must certainly play a part. If it did not, it would follow that humanity's future existence may be good, all things considered, even if none of the future lives were good, or even if they were horrific. The quality of the future lives must be at least part of what determines the goodness of humanity's future existence.

Having noted this, opponents of the stronger welfare aggregation view may instead object that the view is implausibly narrow. The view, they may insist, ignores other important forms of value that also contribute to the goodness of humanity's future existence. That is, on this alternative approach, there are *additional* values (beyond the goodness of the future lives) that partly make humanity's continued existence good. Call views of this type *additional value views*. And let me give a few representative examples.

There are of course many ways of developing this line of thought. Johann Frick (2017), for instance, argues that we have moral reasons to prevent humanity's extinction grounded in the "final" value of humanity. As Frick sees it, humanity—similarly to wonders of nature, great works

[1] See Bostrom (2008).

of art, cultures, etc.—is valuable for its own sake, and appropriately responding to this value, in turn, requires a disposition to ensure its survival. So, he seems to think, the goodness of humanity's continued existence consists at least partly in humanity's "final" value, and partly in the value of the future lives (since, he concedes, the disposition to ensure humanity's survival must be defeasible, as it may be defeated when the lives would be very bad).[2] Similarly, Samuel Scheffler (2018) has argued that our reasons for preventing humanity's extinction are not exhausted by considerations of the goodness of future lives. Specifically, we have reasons deriving from the value of humanity itself—or what he calls "attachment-independent" reasons—to "[try] to ensure that the chain of human generations is extended into the indefinite future under conditions conducive to human flourishing" (Scheffler 2018, p. 103).[3] Here too, then, the idea seems to be that the value of humanity's continued existence consists at least partly in the value of humanity itself, and partly in the goodness, or "flourishing," of the future lives. Finally, Patrick Kaczmarek and Simon Beard (2020) suggest that we have reasons of a rather different kind for preventing humanity's extinction.[4] These have to do, not with the goodness of future lives or our obligations to future generations, but with our obligations to the past. In particular, Kaczmarek and Beard think, by failing to ensure the continued existence of humanity, we would "wrong past people who had anticipated and made grave sacrifices for the benefit of future generations and the long-term future of our species. Only if future generations actually exist, and enjoy lives that are worth living, will these sacrifices be worthwhile" (2020, p. 201). So, while Kaczmarek and Beard focus on "worthwhileness" rather than goodness or value, the basic idea again seems to be that the value of humanity's continued existence consists at least partly in the value of making past generations' sacrifices worthwhile, and partly in the goodness of the future lives (or, in the lives' being "worth living").

Additional value views are then far from outliers. And, in fact—at least unless we have strong utilitarian leanings—I think they would seem the intuitively more attractive view. In contrast to how the welfare aggregation view seems to treat them, future people are not mere "containers" of goodness or well-being, the aggregation of which entirely accounts for the value of their, and indeed humanity's, future existence. While their welfare is of course relevant to, and an important part of, the value of their existence, there seems to be something more to what makes humanity's continued existence especially valuable. Appealing to humanity's "final" value, or the value of the existence of valuers themselves, or the value of realizing the benefits of past generations' sacrifices, thus seems a generally attractive approach.

In this paper, however, I will argue that additional value views are untenable. Despite their initial appeal, these views will inevitably face conflicts between the additional values to which they appeal, and the value of the welfare of future lives. In these cases, it seems the views will imply that their additional values can make up for the lack of value, or even the negative value,

[2] Frick (2017, p. 362). Finneron-Burns similarly interprets Frick's view, writing: "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" (2024, p. 103). Indeed, Frick claims that humanity's value "gives us a reason to ensure, not that humanity grimly soldiers on, but that it survives in a *flourishing* state" (2017, p. 360). At the very least, in any case, given Frick's claims about the defeasibility of his proposed disposition, it seems clear that the *negative* welfare of the future lives partly determines the value of humanity's continued existence, on his view, and so that the view remains an "additional value view." I should also note that, since my arguments below turn primarily on the value of lives that would not be worth living, they will apply equally to Frick's view, regardless of the relevance of flourishing lives.

[3] Scheffler (2018) also offers several "attachment-based" reasons for caring about the fate of future generations. Since my arguments won't concern reasons of this kind, however, I will not focus on them here.

[4] For a convincing critique, see Finneron-Burns (2022).

of the welfare of future lives. If they can avoid this implication, they can do so only by appeal to arbitrary restrictions on when their additional values are indeed valuable. That, at least, is what I will argue here.

1 | AGAINST ADDITIONAL VALUE VIEWS

Begin by considering the following pair of possible outcomes:

A) People continue being born into the indefinite future, but with lives that are all miserable (that is, very bad, not worth living).
B) Extinction.

For present purposes, we can understand extinction to mean the end of humanity, including the end of all presently existing lives as well as the possibility of any future ones. This ending need not be "swift," as it may be in the case of (e.g.) a nuclear war. Alternatively, the end may result from there simply being no additional people who (could) come into existence, while the final generation of people live out the rest of their lives. So, in weighing A and B, we need not take into account the badness of the premature deaths of all existing people. To set terms, we can imagine that, in B, humanity would die out in this way in (say) the year 2200. Moreover, in A, we should suppose there is no chance of the lives ever improving, and so that all future lives would invariably be uniformly miserable.

Faced with a choice between A and B, in any case, which should we choose? Clearly, or so it seems to me, we should choose B. Typically, when considering the badness of human extinction, we are especially distressed by the prospect of no further good lives ever again coming into existence. What's so upsetting, we think, is the great potential that extinction would destroy. Yet when all that's destroyed is an indefinitely long future of miserable lives and suffering, it seems clear that there is not much for us to be upset about. What should upset us, instead, is the prospect of this very bad future. So, it seems clear enough to me, we should prefer B to A.

The welfare aggregation view clearly endorses this choice. After all, if all that's bad about extinction is that it would preclude the existence of future lives worth living, then it will not be bad when the only alternative is one where all future lives are miserable. If there is no possibility where there are future lives worth living, on the welfare aggregation view, then the badness of extinction—which derives entirely from the loss of potential lives worth living—is extinguished. Defenders of the welfare aggregation view are then well-positioned to endorse the intuition that B would be better than A.

Additional value views seem similarly well-positioned—at least, on any plausible precisification of them. While these views admittedly leave open the extent to which their additional values are indeed valuable, it seems clear, given their qualifications regarding the goodness of future lives, that they will not take their additional values to outweigh the badness of generation after generation of miserable lives. If they did, after all, it would amount to claiming that (e.g.) humanity's final value was so great as to outweigh practically any amount of badness in the lives making humanity up. Humanity, in other words, could be constituted by nothing but very bad lives, and yet still be so valuable in itself that it required promotion into the indefinite future. This is implausible. Unless we are committed to a grossly extreme axiology on which

the existence of (say) human beings, or "rational" beings, or fully autonomous beings, or the like, have some type of "absolute" value, or value that outweighs any amount of badness contained in the lives of those beings, we should reject that humanity has such overwhelming value. And so, we should reject that humanity's (putative) additional value could make A better than B.

Additional value views, as I say, may endorse this conclusion. They claim that humanity's continued existence has *some* additional value, outside of the goodness of the future lives, but they've never claimed it has *that* much value. This, though, raises the obvious question: *then how much additional value does it have*? We now have a clear idea of how much value it does not have, but it then seems natural to wonder how much value, if any, it might have. And, here, I'm wary of the potential answers.

In A, the lives would be miserable. And, of course, lives can be bad without being miserable. So let's consider another pair of possible outcomes:

C) People continue being born into the indefinite future, but with lives that are all barely not worth living.
B) Extinction.

In C, the lives would be (uniformly and inevitably) bad, but not miserable. In particular, they would be just barely not worth living. Still, humanity would continue existing. Does this make C better than B?

Again, the welfare aggregation view answers No. Since the value of humanity's continued existence is derived entirely from the goodness of the lives it contains, this view claims, a humanity containing just lives that would not be worth living would not be better than extinction. Extending the view to account for disvalue, it claims that C would be worse than B, given the great amount of badness contained in the lives making humanity up. So, again, the welfare aggregation view gives a straightforward, and not obviously implausible, answer.

What of additional value views? Here, admittedly, I can only speculate. What I imagine, though, is that proponents of these views would want to deny that humanity's continued existence would make C better than B.[5] (As we'll see, little will hang on whether I'm correct about this.[6]) Granted, they may again claim, humanity's continued existence has some additional value, outside of the goodness of the future lives, but this value again fails to outweigh the great amount of badness contained in the very many future lives that would be barely not worth living. In fact, I would hope proponents of these views would claim, this great amount of badness makes C worse than B, and so we should prefer B to C.

[5] Again, this response seems to fit with the views' qualifications regarding the goodness of the future lives—that they must be "flourishing" or at least "worth living." The response is also suggested by at least some of the claims of the views' proponents. Frick, for instance, briefly discusses a related case, where just the lives of our more immediate descendants would not be worth living (unlike in C, where all future lives would not be worth living). And, even here, he suggests that his proposed moral reason to ensure humanity's survival "would presumably be outweighed or cancelled" (Frick 2017, p. 362). Similarly, Scheffler notes that we are not "concerned solely with humanity's bare survival. Rather, we want future generations to survive under conditions conducive to their flourishing" (2018, p. 60). And while he does not elaborate much on these conditions conducive to flourishing, it seems plausible that they would not be present in C.

[6] As I mention below, the arguments I'm about to offer generalize. So even if proponents of additional value views deny that B is better than C, we can simply replace C with whatever relevantly worse outcome they choose (even if this is, say, A).

Here, though, are the beginnings of a puzzle. To see it, compare next another pair of possible outcomes:

D) People continue being born into the indefinite future, with lives that are all barely worth living.
B) Extinction.

Presumably, all would agree that D would be better than B.[7,8] The lives in D may be (uniformly and inevitably) barely worth living, but they remain worth living, and the (putative) additional value of humanity's continued existence would be secured. So, we would have nothing telling in favor of B, and at least one—and perhaps two, if we accept an additional value view—strong consideration(s) in favor of D.

Now, though, we may imagine a continuum of possible outcomes, specifically between C and D. In C, the lives would be uniformly barely not worth living, and in D, the lives would be uniformly barely worth living. The continuum I have in mind is made up of the possible outcomes in which we "flip" the quality of a given life in C or D to look like a life in the other, one by one, until we move entirely from C to D or vice versa. That is, if we start from C, we can imagine a nearby possible outcome in which all the lives, minus one, remain barely not worth living, and the one other life is barely worth living. We can then imagine a nearby possible outcome in which all the lives, minus two, remain barely not worth living, and the two other lives are barely worth living. We can then repeat this process until we end up with a possible outcome where all the lives are barely worth living—or, until we end up with D.

For each of these possible outcomes, we can compare it with extinction. Would it be better, worse, or equally as good as, extinction? The welfare aggregation view will in each case have a straightforward answer: if the balance of welfare is positive, the outcome is better than extinction; if negative, the outcome is worse than extinction; and if neutral, the outcome is equally as good as extinction. Intuitively, these answers are at least not obviously implausible: if there are more good lives than bad, then humanity's continued existence is better than extinction, and if not, not. Whether we should accept these verdicts depends on whether there are any more plausible ones available.

Do additional value views fare better? It is unclear. After all, on additional value views, matters are not so simple. Presumably, I have said, the views will imply that D—where all the lives are barely worth living—would be better than extinction. They will also imply, I've suggested, that

[7] Admittedly, those who accept "critical level" views may disagree. On these views, there is some positive level of welfare—typically higher than the level at which a life is barely worth living—at which lives begin to contribute to an outcome's goodness. However, there are a number of strong reasons to reject these views. For one, notice that the implication that D would not be better than B is itself a reason to reject the views. And for another, more generally, the views have been shown to imply various "sadistic conclusions," which (as the name indicates) are extremely counterintuitive. See Arrhenius (2000, pp. 255–256) and Huemer (2008, pp. 912–913).

[8] When discussing optimal population sizes, Frick claims (as we've seen) that there is at least some extent to which "responding appropriately to humanity's final value gives us a reason to ensure, not that humanity grimly soldiers on, but that it survives in a *flourishing* state" (2017, p. 360). Of course, it's not entirely clear what counts as "flourishing" or "grimly soldiering on" in this context, and so it's not entirely clear whether humanity's final value gives us reason to choose D over B, on Frick's view. However, as we've also seen, Frick only explicitly claims his proposed moral reason to ensure humanity's survival would be "outweighed or cancelled" in circumstances where the lives would be not worth living (see fn. 5), so I assume for present purposes it would not be outweighed or cancelled in this case. For more on this and related topics, see Finneron-Burns (2024). Thanks to an anonymous referee for pressing me on this.

C—where all the lives are barely not worth living—would be worse than extinction. At some point on our continuum from D to C, therefore, the views will imply that we have gone from an outcome that would be better than extinction, to an outcome that would not be better than extinction. But it is unclear where exactly this point might be. Presumably—a qualifier I am, again, forced to use—the point would be further down the continuum than it is on the welfare aggregation view, or further down than the outcome in which the total welfare was neutral. After all, if humanity's continued existence has additional value, outside the goodness of the future lives, then an outcome in which the total welfare was neutral yet humanity continued existing would presumably, in virtue of that additional value, be better than extinction. Thus, if that's right, additional value views must imply that at some point between this outcome, where the total welfare was neutral, and C, where the lives are all barely not worth living, we move from an outcome that is better than extinction, to one that is not better than extinction.

Consider the first step on this part of the continuum. We move from the outcome where the balance of lives that would be barely worth living and lives that would be barely not worth living was neutral, to an outcome where just one of these lives that would be barely worth living is "flipped," and would now be barely not worth living, and all else is equal. That is, we move to an outcome where half of the lives, minus one, would be barely worth living, and half of the lives, plus one, would be barely not worth living. Call this the *barely-welfare-negative outcome*. Do additional value views imply that this second outcome, where the total welfare would be negative—albeit barely so—would also be better than extinction? Again, of course, it is unclear. Regardless of the answer, though, additional value views seem to face trouble. Let me explain.

Suppose first additional value views imply that this barely-welfare-negative outcome would not be better than extinction. In virtue of the negative balance of total welfare, proponents of the views might claim, the outcome cannot be better than extinction. As soon as the total welfare becomes negative, the possibility of the outcome's being better than extinction is ruled out. This response, they may further claim, is in line with their stated caveat that humanity may have their putative additional value only if the future lives it contains are "flourishing" or, at least, "worth living."[9] In this case, additional value views would be not all that dissimilar to the welfare aggregation view: the latter implies that the outcome where the total welfare would be neutral would not be better than extinction; the former, that while this outcome would be better than extinction, the one where we "flip" the quality of just one of the lives would not be better than extinction.

This is troubling for additional value views, for at least two reasons. First, it implies that, for all the talk of humanity's "final" or "ultimate" value, this value really only amounts to the difference between one person's life being barely not worth living rather than barely worth living. That, it seems, is hardly all that significant, or at least hardly as significant as proponents of these views seem to think. Second, and relatedly, this response implies that additional value views are in fact *very* similar to the welfare aggregation view. In particular, the views differ only in whether they attribute a relatively tiny amount of value to humanity's continued existence, outside of the

[9] As we've seen, proponents of these views often make claims that suggest something like this response. And, in fact, this response may seem the one most faithful to their views. It must be, for Scheffler, that future generations live "under conditions conducive to human flourishing" (2018, p. 102); or for Kaczmarek and Beard, that they "enjoy lives that are worth living" (2020, p. 201); or for Frick, that they do not have lives that are "not worth living" (2017, p. 362). It seems reasonable to think, then, that non-negative total welfare is a condition that must be met in order for humanity to have the putative additional values, at least on some versions of these additional value views. Since these versions of additional value views attribute additional value to humanity only on the condition that those making humanity up have (at least) non-negative total welfare, we might call them *conditional* value views. As we are about to see, views of this kind face major worries.

goodness of the future lives it would contain. The value of the future lives would almost always outweigh any additional value that humanity's continued existence might have—indeed, the additional value would be the deciding factor against extinction *only* when the total welfare of the future lives would be precisely neutral. In that case, we may begin to wonder why we need to bother positing this additional value in the first place. Its existence is in question, and its weight, if it does exist, is relatively tiny. So, this line of response ultimately seems to undermine additional value views, even if it secures their constraint that for humanity to have the additional values, the future lives must be "flourishing" or "worth living."

Suppose next, then, that additional value views imply that the barely-welfare-negative outcome would be better than extinction. We should then, of course, ask: at what point on the continuum does this answer change? That is, we should ask: which outcome on the continuum between the barely-welfare-negative outcome and C (where all the lives would be barely not worth living) is such that it—in contrast to the barely-welfare-negative outcome, but in alignment with C—would not be better than extinction? And here, it is difficult to see how a justified answer might be offered. After all, the claim that the relevant outcome is the barely-welfare-negative outcome is at least justified by the constraint that for humanity to have the additional value, the future lives must be "worth living"—or, at least, have a total welfare that is not negative. Once we abandon this constraint, though, we are left to weigh the weight of the putative additional value against the weight of the total well-being—or really, in this case, the total *ill*-being—of the future lives. The weight we assign to the additional value will determine at what point on the continuum we reach an outcome that, despite ensuring humanity's continued existence, we should not prefer to extinction. Obviously, the weight we should assign to the additional value is just the extent of the value it actually has. We know (or, at least, are assuming) that this amount is not great enough to outweigh the badness of there being indefinitely many future lives that would all be barely not worth living; currently, the suggestion is that it would be great enough to outweigh the net badness of there being indefinitely many future lives, half of which, minus one, would be barely worth living, and the other half of which, plus one, would be barely not worth living. The question, then, is how many of these lives we must subtract from the one half and add to the other, until the additional value no longer outweighs the net badness of the lives. And what reason could we have to adopt any particular answer? Here, there seems to be no principled response available.

Abandoning this apparently unworkable approach, proponents of additional value views might retreat to a haven that is popular among those discovering a deep difficulty in weighing between competing values: they might appeal to a form of *incommensurability*. Note that I have recently not spoken of the implications of additional value views in terms of whether an outcome would be *worse* than extinction; rather, I've spoken in terms of whether an outcome would be *not better* than extinction. This is important, because it may be possible for an outcome to be neither better than, nor worse than, nor precisely equally as good as extinction. Indeed, it may be possible, given the qualitative differences between the outcomes, that another value relation holds between them. While there a number of ways of developing this line of thought, I'll follow Parfit (2016) in calling the relevant relation "imprecise equality."[10] (Similar notions here include "parity," "rough equality," and "the negative intransitivity of *better than*."[11]) Two things, x and y, are *imprecisely*

[10] More generally, Parfit appeals to the notion of "evaluative imprecision:" he thinks things can be imprecisely better than, imprecisely worse than, or imprecisely equal to, others. In these cases, the two things are cardinally comparable, but not by some number of units on a relevant scale. See especially Chang (2016). See also Fanciullo (2019). Since nothing below will hang on these other "imprecise" relations, I focus just on "imprecise equality" for simplicity.

[11] See e.g. Chang (2002), Chang (2016), Griffin (1986), Hare (2013), and Hurka (1993).

equally good if, while *x* is neither better nor worse than *y*, there could be some third thing, *z*, that is neither better nor worse than *x*, but that *is* better or worse than *y*.[12] Take, for instance, the question of whether Einstein or Bach was the greater genius.[13] Given the qualitative differences between scientists and composers, it may seem plausible to think, it simply is not true that either was a greater genius than the other. Still, you might think, it also isn't the case that their geniuses were precisely equally great. After all, if they were, then (say) a single additional paper from Einstein could have tipped the scales, and made him a greater genius than Bach. And this seems wrong. Instead, it seems that the qualitative differences in what constitute their geniuses make them such that neither was a greater genius than the other, yet their geniuses were also not precisely equally great.[14] In this case, we can say that their geniuses were *imprecisely* equally great: Einstein was neither a greater nor lesser genius than Bach, yet if Einstein had written another paper that contributed to his genius, while he still would have been neither a greater nor lesser genius than Bach, he *would* have been a (slightly) greater genius than he actually was.

Proponents of additional value views might adopt a similar strategy. Earlier, I suggested that if these views imply that the barely-welfare-negative outcome would not be better than extinction, then the views' implications would hardly be different from those of the welfare aggregation view. This is because, just as the welfare aggregation view implies the barely-welfare-negative outcome would not be better than extinction, so additional value views imply this. The welfare aggregation view, though, does not just imply the barely-welfare-negative outcome would not be better than extinction, but it implies that the barely-welfare-negative outcome would be *worse* than extinction. Because of this, the view implies that we should choose extinction over this barely-welfare-negative outcome, because extinction would be better. Additional value views, on the other hand, need not imply this, because they may take the two outcomes to be (e.g.) imprecisely equally good. This response may be laid out as follows.

As the example of Einstein and Bach illustrates, imprecise equality typically arises from significant qualitative differences in value. If we are weighing the difference between ten seconds of a pleasurable sensation and two five-second experiences of that very same sensation, for instance, it will hardly seem plausible to claim the two are imprecisely equally good. In the case we're considering, though, proponents of additional value views might claim there clearly is a qualitative difference in the values being weighed: on the one hand, the value and disvalue of individual lives; on the other hand, the additional value involved in humanity's continued existence, such as its "final" value. When comparing a combination of these (dis)values to extinction, then, it might be that the outcome containing the combination is merely imprecisely equal to an outcome containing no part of the combination.[15] Thus, for instance, it might be that the barely-welfare-negative outcome, containing the slight net badness of the future lives as well as the "additional value" realized by humanity's continued existence, would be neither better nor worse than, nor precisely equally as good as, extinction. In addition, it might be that the adjacent outcome on

[12] Parfit (2016, p. 115).

[13] This example is due to Parfit (2016, pp. 113–114), though, as he notes, examples of this more general kind are attributable to Chang (2002).

[14] While this may lead one to think the geniuses of scientists and composers simply cannot be compared, notice that Einstein was clearly a greater genius than many poor composers, and Bach was clearly a greater genius than many poor scientists.

[15] To extend the above analogy about pleasure, the idea might be that whereas ten seconds of a pleasurable sensation and two five-second experiences of that same sensation are precisely equally good, ten seconds of the pleasurable sensation and eleven seconds of a correspondingly painful sensation may be imprecisely equal to experiencing no sensations at all.

our continuum—where there would be indefinitely many future lives, half of which, minus two, would be barely worth living, and the other half of which, plus two, would be barely not worth living—would be neither better nor worse than, nor precisely equally as good as, extinction. But still, it might also be that one of these two adjacent outcomes *would* be better than the other: the barely-welfare-negative outcome would be better, because it would contain one extra life that was barely worth living, and one less life that was barely not worth living. In this way, the barely-welfare-negative outcome and extinction may be imprecisely equally good: neither outcome would be better or worse than, or precisely equally as good as, the other, yet there would be a third thing—the outcome adjacent to the barely-welfare-negative outcome—which despite being neither better nor worse than, nor precisely equally as good as, extinction, *would* be worse than the barely-welfare-negative outcome. In this way, proponents of additional value views may resist my claim that their views are hardly different from the welfare aggregation view: whereas the latter implies we should choose extinction over the barely-welfare negative outcome, additional value views imply nothing of the sort, since the two outcomes are simply imprecisely equally good.

This response is certainly coherent, and some may even find it plausible. Intuitively, even I must admit, it is difficult to definitively judge that extinction would be better or worse than, or precisely equally as good as, these outcomes that would be barely negative in total welfare. So why not appeal to imprecise equality, and avoid the issue altogether? Because I think we can do better. And, indeed, I think we must, because the appeal to imprecise equality leads to a familiar problem. Specifically, I've suggested that if additional value views imply that the barely-welfare-negative outcome would be better than extinction, then they will lack a justified answer as to at what point on the continuum between the barely-welfare-negative outcome and C (where all the lives would be barely not worth living) we would reach an outcome that—in contrast to the barely-welfare-negative outcome, but in alignment with C—would not be better than extinction. On the current response, while the barely-welfare-negative outcome would not be better than extinction, it would also not be worse than or precisely equally as good as extinction, either. But here is the problem. Presumably, I have said, proponents of additional value views will want their views to imply that C would be worse than extinction—not imprecisely equal, but worse. In that case, we can again ask: at what point on the continuum between the barely-welfare-negative outcome and C would we reach an outcome that—in contrast to the barely-welfare-negative outcome, but in alignment with C—would be, not imprecisely equal to, but *worse* than extinction. And, again, it is difficult to see how any potential answer might be justified.

One possible answer may seem familiar. Perhaps the point at which the outcomes and extinction are no longer imprecisely equally good is just the point where we move from the outcome adjacent to C, to C. That is, perhaps C is the first point on the continuum between the barely-welfare-negative outcome and C at which the outcome is, not imprecisely equal to, but worse than extinction. And, perhaps, this is because there is no longer a combination of goodness and badness in the future lives that would exist: there is no single person who would have a life that was (barely) worth living. Because of this, C would indeed be worse than extinction, while the adjacent outcome—containing an indefinite number of future lives, just one of which would be barely worth living, and the rest of which would be barely not worth living—would be imprecisely equal to extinction. That, at least, is the only principled response I can think of on behalf of additional value views.

But I also think we should reject this response. Presumably, if we think C would be worse than extinction, changing *just one* of these indefinitely many future lives to be barely worth living rather

than barely not worth living will not, or at least should not, radically change our assessment. The adjacent outcome, I would think, would also be worse than extinction—not imprecisely equal, but worse. Perhaps others will disagree. As I see it, though, this would require a staunch and, frankly, unlikely defense. Our default position should be to reject this response. Thus, it seems the problem for additional value views that appeal to imprecise equality (or "rough equality," "parity," etc.) remains. They must identify the first outcome on the continuum between the barely-welfare-negative outcome and C that—in contrast to the barely-welfare-negative outcome, but in alignment with C—would be, not imprecisely equal to, but *worse* than extinction. And, again, it is difficult to see how any potential answer might be justified.

Let me briefly sum up the ground we've covered. I've first suggested that, on both additional value views and the welfare aggregation view, C would be worse than extinction, and D would be better than extinction. I then introduced the idea of an outcome, near the middle of a spectrum of outcomes between C and D, that would be just barely negative in total welfare: in this barely-welfare-negative outcome, there would be indefinitely many future lives, half of which, minus one, would be barely worth living, and the other half of which, plus one, would be barely not worth living. I've then asked whether this outcome would be better than extinction. On the welfare aggregation view, I've said, it would not be—indeed, it would be worse than extinction, since the total welfare would be negative. I've then canvassed the potential answers to this question on behalf of additional value views. Obviously, the barely-welfare-negative outcome must either be better than extinction, or not be better than extinction. If they claim it would be better, I've suggested, then they must identify the point on the continuum between this outcome and C (which they claim would be worse than extinction) at which their answer changes, or the first outcome that would not be better than extinction. I've then suggested that the choice of any particular outcome here would be arbitrary. On the other hand, if additional value views claim the barely-welfare-negative outcome would not be better than extinction, they have three main options. First, they may claim it would be worse than extinction, in which case they will hardly differ from the welfare aggregation view. Second, they may claim it would be precisely equally as good as extinction, in which case they will, again, hardly differ from the welfare aggregation view. And third, they may claim that the two outcomes are imprecisely equally good: neither would be better or worse, or precisely equally as good as, the other. In this case, I've claimed, the views run into a familiar problem: they must identify the point on the continuum between the barely-welfare-negative outcome and C (which they claim would be *worse* than extinction) at which their answer changes, or the first outcome that would be, not imprecisely equal to, but worse than extinction. And here again, I've suggested, the choice of any particular outcome would be arbitrary.

Now, much of what I've said here has been rather speculative: since additional value views tell us little about the extent to which humanity's continued existence has additional value, I have had to offer what I take to be plausible responses on their behalf when comparing outcomes. It is worth noting, though, that my points here actually generalize. So suppose, for instance, that proponents of additional value views object to my suggestion on their behalf that C would be worse than extinction. Perhaps they think it would be better than extinction, or that the two would be imprecisely equally good. Presumably, though, given their caveats about the future lives being "flourishing" or, at least, not very bad ones, there must be some negative level of welfare at which, they would say, an indefinite number of future lives at this level would be worse than extinction. And, in that case, we can run my arguments in just the same way, simply replacing "C" with this outcome, and "D" with an outcome where the lives would be correspondingly

high in welfare (so as to mirror the negative level of welfare in the replacement for "C"). In this case, the arguments—and, presumably, our intuitions—would apply just the same. Regardless of whether proponents of additional value views have the starting intuitions I've ascribed to them, then, the structure of the arguments here will remain unaffected.

A final lingering objection here concerns my charges of arbitrariness. I've said that if additional value views claim either that the barely-welfare-negative outcome would be better than extinction, or that the two would be imprecisely equally good, they will lack any principled way of identifying the first outcome on the continuum between the barely-welfare-negative outcome and C that would be—like C, but unlike the barely-welfare-negative outcome—worse than extinction. At this point, however, it may be wondered why this is even a problem, or why the views would even need to offer anything like the kind of justification I'm calling for. After all, even if we lack a principled way of identifying the relevant outcome, its identity—or, which outcome it actually is—will presumably be determined just by the axiological facts, or the ultimate facts about the combinations of value in the different outcomes and their relations to the outcome in which human beings would become extinct. Thus, the fact (if it is one) that additional value views have no principled way of identifying the relevant outcome does not imply that it does not exist, or that additional value views must be false. Instead, it implies just that the relevant axiological facts are, to us, at least at the moment, opaque.

That seems right, I think. At the very least, I don't take myself to have shown that additional value views must be false. And, admittedly, it could turn out that humanity's continued existence does have the type of additional value these views claim it to have. What I take myself to have shown here, however, is that what at first seems an intuitively very attractive idea—that there is something additionally valuable about humanity, outside of the welfare of the lives making it up—is upon closer inspection axiologically suspect. If this idea, as I've characterized it here, turns out to be true, then we will seem to lack any principled way of identifying its implications. Indeed, we will lack any principled way of determining the weight of the additional value it purports humanity's continued existence to have, or the extent of this value. This, I think, should make us suspicious of this idea, despite its clear initial plausibility. In fact, it seems, we should be especially suspicious, given we have an alternative view, the welfare aggregation view, whose implications are in each case entirely straightforward, and which focuses on the source of value that in each case seems most important of all: welfare.

2 | SUMMING UP

In this paper, I've offered some reasons for being wary of additional value views. If they are true, then they will be true despite our lacking any principled way of determining their implications, or indeed the weight of the additional value they attribute to humanity's continued existence. To be sure, there is a good deal of initial plausibility in the idea that there's *something* about humanity's continued existence—whether it be humanity's final value, or the value of the existence of valuers themselves, or the value of past generations' sacrifices being made worthwhile—that makes it additionally valuable, outside of the value of the lives making it up. The problem, though, is that these putative additional values will at some point have to be weighed against the disvalue of lives that are not worth living. The lives will either have to be preferred as a means to ensuring the putative additional value of humanity's continued existence, or make things such that the outcome in which they would exist would be neither better nor worse than, nor equally as good

as, extinction. Either way, we will seem to lack justification for any particular weight we assign to the putative additional value. It seems, therefore, that additional value views are on shaky ground. Without further justification, they themselves may face extinction.

ACKNOWLEDGEMENTS

Many thanks to an anonymous referee for helpful comments.

ORCID

James Fanciullo https://orcid.org/0000-0002-9397-1951

REFERENCES

Arrhenius, G. (2000). An Impossibility Theorem for Welfarist Axiologies. *Economics and Philosophy*, 16, 247–266. https://doi.org/10.1017/S0266267100000249

Bostrom, N. (2008). Why I Want to Be a Posthuman When I Grow Up. In B. Gordijn & R. Chadwick (Eds.), *Medical Enhancement and Posthumanity* (pp. 107–137). Dordrecht: Springer. https://doi.org/10.1007/978-1-4020-8852-0_8

Chang, R. (2002). The Possibility of Parity. *Ethics*, 112, 659–688. https://doi.org/10.1086/339673

—. (2016). Parity, Imprecise Comparability and the Repugnant Conclusion. *Theoria*, 82, 182–214. doi: 10.1111/theo.12096

Fanciullo, J. (2019). Imprecise Lexical Superiority and the (Slightly Less) Repugnant Conclusion. *Philosophical Studies*, 176, 2103–2117. https://doi.org/10.1007/s11098-018-1117-4

Finneron-Burns, E. (2022). Human Extinction and Moral Worthwhileness. *Utilitas*, 34, 105–112. doi: 10.1017/S095382082100039X

—. (2024). 'Humanity': Constitution, Value, and Extinction. *The Monist*, 107, 99–108. doi: 10.1093/monist/onae001

Frick, J. (2017). On the Survival of Humanity. *Canadian Journal of Philosophy*, 47, 344–367. doi: 10.1080/00455091.2017.1301764

Griffin, J. (1986). *Well-Being: Its Meaning and Measurement*. Oxford: Oxford University Press.

Hare, C. (2013). *The Limits of Kindness*. Oxford: Oxford University Press.

Huemer, M. (2008). In Defence of Repugnance. *Mind*, 117, 899–933. https://doi.org/10.1093/mind/fzn079

Hurka, T. (1993). *Perfectionism*. Oxford: Oxford University Press.

Kaczmarek, P. & Beard, S. (2020). Human Extinction and Our Obligations to the Past. *Utilitas*, 32, 199–208. https://doi.org/10.1017/S0953820819000451

Parfit, D. (2016). Can We Avoid the Repugnant Conclusion? *Theoria*, 82, 110–127. doi: 10.1111/theo.12097

Scheffler, S. (2018). *Why Worry About Future Generations?* Oxford: Oxford University Press.

How to cite this article: Fanciullo, J. (2024) Why prevent human extinction? *Philosophy and Phenomenological Research*, 1–13. https://doi.org/10.1111/phpr.13066