Philosophical Studies
https://doi.org/10.1007/s11098-024-02189-5

Group prioritarianism: why AI should not replace humanity

Frank Hong[1]

Accepted: 26 June 2024

Abstract
If a future AI system can enjoy far more well-being than a human per resource, what would be the best way to allocate resources between these future AI and our future descendants? It is obvious that on total utilitarianism, one should give everything to the AI. However, it turns out that every Welfarist axiology on the market also gives this same recommendation, at least if we assume consequentialism. Without resorting to non-consequentialist normative theories that suggest that we ought not always create the world with the most *value*, or non-welfarist theories that tell us that the best world may not be the world with the most *welfare*, I propose a new theory that justifies giving some resources to humanity in the face of overwhelming AI well-being. I call this new theory, "Group Prioritarianism".

Keywords AI · Population ethics · Utility monsters · Prioritarianism · Super-beneficiaries · AI Safety · AI-Wellbeing

1 Introduction

In 1974, Robert Nozick famously introduced the idea of the "utility monster", a being that is so efficient at converting resources into utility, that the best way to maximize utility in the world is to give all resources to this one monster at the cost of all of humanity. Nozick goes on to say that the mere possibility of the utility monster refutes Utilitarian theories, which imply that one ought to sacrifice all humanity into the maws of this utility monster. Such a consequence is taken to be so outlandish that any theory that suggests that humanity ought to be sacrificed to the monster is, in Nozick's words, "embarassed" (Nozick, 1974, p. 41).

In this paper, we will take as a fixed point that Nozick is right in this case: it really is wrong to sacrifice all humanity to the monster, even if doing so increases

✉ Frank Hong
franksho@usc.edu

[1] Hong Kong University, Hong Kong, China

Published online: 13 July 2024

total well-being.[1] But even if we take this as a fixed-point, many important questions are left unanswered. For example, should we create these utility monsters if we had the option? And if these monsters do come to exist, and if we ought not to give everything to the monsters, then *how much should* we give to these monsters? And if Utilitarianism cannot accommodate our intuitions about distributive welfare in this case, then what kind of normative theory can?

These questions have gained a new urgency as the utility monster no longer exists as a mere thought experiment designed for the sole purpose of refuting Utilitarianism. It is possible that in the future, these utility monsters might actually come to exist. Bostrom and Shulman (2020) argue that these utility monsters may indeed be realized as artifical agents, or digital minds, who can be cheap to produce and who can possibly experience hundreds of subjective years of life in a short period of "wall-clock time".[2] Bostrom and Shulman call these artificial agents, "superbeneficiaries" (henceforth, "supers"). This raises questions about whether we should indeed create such supers if we can; and if so, how much of our resources should we give them? To be told that Utilitarianism is false is not enough guidance.

In this paper, I will be developing a novel axiology—one that is independently motivated by our intuitions regarding fairness and diversity—to answer these questions and to vindicate the judgment that we should not replace all of humanity with supers. Moreover, our axiology will be welfarist (i.e. it will only take *well-being* as a value). And finally, we will be assuming *consequentialism*, which tells us that facts about what acts we ought to perform depend only on how good the consequences of those acts are.

Here I want to stress that in making all these assumptions, I am not dismissing other normative theories (like deontology) or other axiologies (like pluralist theories of value) as being obviously false. In fact, I think that these views are not obviously false, and that we may indeed have non-consequentialist reasons to not replace humanity,[3] And I think that there may be other values besides well-being that would be lost if we replaced humanity. However, I think it would be a shame if Team Humanity is open only to deontologists and pluralists. I think all humans should be on Team Humanity and resist the idea that we should be replaced by supers. Indeed, if we have a bunch of welfarist population ethicists telling us that it is *obligatory* to

[1] Instead of talking about "utility" I will talk about well-being for two reasons. The first is that the term "utility" is often used not as units of real well-being, but a function that is used to represent the structure of one's preferences. Secondly, whereas "utility" is often associated with preference-satisfaction, I want to remain neutral about whether well-being is determined by the satisfaction of one's preferences, or whether it is determined by some pleasurable mental state, or even whether it is determined by the possession of a list of objective goods.

[2] Bostrom and Shulman actually give eight different reasons why one might think an AI can be a superbeneficiary apart from their ability to experience more subjective years of life given a short period of "wall-clock time". For example, perhaps we have reason to think that these digital minds have a greater "hedonic range" such that they can experience states of pleasure that are physically impossible for us to instantiate. The details need not detain us here, but they are worth mentioning because not all would agree that a supers ability to experience more subjective years of life would entail that they are able to enjoy anymore well-being than a human (see Mogensen (2023)).

[3] See (Nebel, 2021; Cohen, 2012) and Scheffler (2018) Scheffler (2018) for broadly non-consequentialist reasons to prevent things of value from being replaced by other things of even greater value.

replace humanity with supers, then this could constitute an existential risk to humanity. So if one likes humanity, it would be better to have welfarists fighting with us, rather than against us.

Unfortunately, the survival of humanity is particularly hard to justify within a welfarist framework when we are faced with the possibility of creating a group of supers. I will develop this challenge in greater depth in Sect. 2. Then, I will take this challenge head-on in Sect. 3 where I motivate and develop our welfarist theory. Finally, in Sect. 4, I will discuss objections to the view. Ultimately, my goal is not to argue that the best way to justify human survival is on welfarist grounds, or even that our particular welfarist theory is the most plausible out of all possible welfarist theories. Rather, my goal is to explore what a welfarist justification for human survival would look like, and to explore its motivations and implications.

2 Resisting replacement

First, we'll assume that supers require less resources to be sustained compared to a human, and so the same amount of resources used to sustain a human life can be used to sustain even *more* super lives. Secondly, let's assume that these supers can enjoy more total well-being than a human possibly can (either because they can experience *more* subjective years of bliss within one objective year of "wall-clock time", or because they can instantiate states of well-being that are physically impossibly for humans to instantiate). To make things more concrete, let us also assume that for every unit of resources that could be used to sustain a happy human life, we can use that same amount of resources to sustain 10 super lives, each with 100 times more total well-being than the happy human life.[4]

So, with these assumptions on the table, we can now consider the following principle:

> **Beneficent replacement**:
> For any population X, if there is a happy human, h, in population X, then a population Y, that is otherwise like X except h is replaced with 10 supers with 100 times more well-being than h, is better than X.

Imagine a future population X_1 where all the universe's resources is devoted to supporting 100 humans (a slight underestimate), and imagine a population X_{100} where instead all the universe's resources are given to ten times as many supers. One can

[4] I should note that, in giving these assumptions, I am not saying they are *obviously* true. It is an open empirical question whether AI are even capable of well-being (see (Goldstein and Kirk-Giannini, 2023) for discussion). And even if so, it may be difficult to ascertain exactly *how* much more well being an AI actually enjoys (see (Fischer and Sebo, 2024)). In reality, our decision to distribute resources to these potential supers is a decision under uncertainty. A fully adequate account of what we should do, therefore, should take in account our uncertainty of whether AI really are capable of well-being. However, for the purposes of this paper, we will assume this uncertainty away. I think it instructive to see how things go when we are not uncertain. Furthermore, I think it helpful to see whether human survival can be justified in the case where we are *certain* that AI are super-beneficiaries. If it turns out that it's impossible for AI to be super-beneficiaries, that would make justifying human survival substantially easier. But what I want to explore in this paper is whether human survival can be justified when things are not so easy.

construct a sequence of populations between X_1 and X_{100} such that each member of the sequence has one less human and 10 more supers than the last. Given **Beneficent Replacement**, each member of the sequence is better than the last, and given the transitivity of the "better than" relation, the last member of the sequence is the best.

Since **Beneficent Replacement** implies that a world exclusively of supers is better than any world with a mix of humans and supers, any axiology that does not imply that a world exclusively of supers is the best must deny **Beneficent Replacement**.

Unfortunately, almost all axiologies imply **Beneficent Replacement**. It is easy to see, for example, that both totalism and averageism imply **Beneficent Replacement** since each replacement increases both total and average well-being.

Similar things can be said about Prioritarianism. According to Prioritarianism, the overall good of a population is determined by a *weighted* sum of each individual's well-being, where the well-being of the worst off "counts for more" than the well-being of the best off. More formally, the overall good of a population is determined by first applying a strictly increasing and concave function on each individual's well-being levels, and then summing them up. A common example of a strictly increasing and concave function is the square root function, and so a typical Prioritarian view would say that overall good is determined by:

$$O(X) = \sum_{i=1}^{n} \sqrt{w(x_i)}$$

where $O(X)$ is the overall good of population X, n is the number of individuals in X, and $w(x_i)$ is the level of well-being of individual x_i in X.

It can easily be seen that even on the Prioritarian view, *replacing* an individual with a much happier individual will increase the overall good, even though the increase is not proportional to the increase in total well-being. In other words, replacing everyone with 10 times more people with 100 times more well-being may not make things *1000 times* better overall, but it will still make things *better* overall.

Even *ad hoc* "Prioritarian" views that say that there is only value in increasing the well-being of an individual up to the point of maximum human level well-being, after which there is no value in any more additional well-being, would *still* imply **Beneficient Replacement**. This is because the sheer fact that you can replace 1 happy human with 10 supers who are at *least* as happy as the human will always increase overall good.

I take it that Totalism, Averageism, and Prioritarianism are some of the most popular and plausible axiologies. Many more besides these that vindicate **Beneficent Replacement** will also get you the result that a world exclusively filled with supers is better than any mix of supers and humans. For example, there are critical level utilitarians who argue that there is a *threshold* of well-being above having a life barely worth living such that anyone who has well-being below that threshold contributes negatively to the overall good. These critical level utilitarians would also accept **Beneficent Replacement** so long as the critical threshold isn't far above the welfare of the best supers (but if that were the case, the critical level utilitarians should advocate for extinction).

Perhaps one can resist **Beneficent Replacement** by adopting a kind of Person-Affecting View. Person-Affecting views are united by a common emphasis on the

Table 1 Conserve or Exploit

	Present population	Future population A	Future population B
Conserve	100	–	100
Exploit	110	50	–

value of “making people happy" over “making happy people" (Narveson, 1973). There are two broad categories of Person-Affecting Views. Some views are *strong* in the sense that they say that a world w' is better/worse than a world w *only if* there is some person who is better/worse off in w' than in w. On a strong view, **Beneficient Replacement** would be false because there *is no* individual in the replacement world who would be better off.[5]

Other views are *weak* in the sense that they say that a world w' is better/worse than a world w *in one important “Person-Affecting" dimension only if* there is a person who is better/worse off in w' than in w.[6] However, on this view, a world w' can still be better than w *overall* even if it is not better in this Person-Affecting Dimension.

The Weak Person-Affecting view is motivated in part by Parfit’s famous “Non-Identity Case" featuring the choice about whether to exploit the environment (Parfit, 1984, pg. 362).

If one exploits, then presently existing people will benefit, but future people will enjoy only a very mediocre life. But if one conserves, then a different set of future people will come to existence, and they will be very well off. The decision is illustrated in the Table 1 below:

Intuitively, one ought to conserve rather than exploit. However, on *Strong* Person-Affecting views, exploiting is not worse than conserving because no one is harmed if we exploit. *Weak* Person-Affecting views, however, do not have this strange consequence. They only have the consequence that the world would not be better off in *one*, albeit important, normative dimension if we choose to exploit rather than conserve. So it’s compatible with Weak Person-Affecting Views to say that conserving is better than exploiting since the world in which we conserve may be better off in other important normative dimensions (e.g. in the dimension of total and average well-being).

[5] Here I am assuming the **Incomparability of Non-Existence** which states that, for any person x with positive well-being in world w, x is neither better off, worse off, nor equally well off in a world w' where x does not exist in world w.

[6] See (Ross, 2015) for an example of such a view.

However, for precisely the reason that Weak Person-Affecting views is compatible with the judgment that we should choose to conserve rather than to exploit, Weak Person-Affecting views are also compatible with **Beneficent Replacement**. They can at most say that each beneficent replacement fails to make the world better in *one* dimension, but they fail to say that each beneficent replacement won't make the world better overall.

So if we want to resist **Beneficent Replacement** by adopting a Person-Affecting view, we need to opt for a Strong Person-Affecting view. On a Strong Person-Affecting View, **Beneficent Replacement** is false because there is no person who is made better off with each replacement.

In fact, Strong Person-Affecting Views would say that each of the X_i are *incomparable* to each other. To see why each of the X_i must be incomparable to each other, as opposed to being equally as good, consider, for example, X_1 and X_2. Suppose that "Super Siri" is a super that exists in X_2 but does not exist in X_1. Now consider X_2+, which is exactly like X_2 except that "Super Siri" has 100 times more well-being. Any plausible Strong Person-Affecting view would at least say that X_2+ is better than X_2 since there *is* a person in X_2+ that is better off than in X_1, and no one else is made worse off. There is only a pareto improvement in well-being from X_2+ and X_2. But if X_2+ is better than X_2, and X_2 is equally as good as X_1, then X_2+ must be better than X_1, which contradicts Strong Person-Affecting views since no one is better off in X_2+ than in X_1. So if one wants to adopt a Strong Person-Affecting View in this context, one must adopt a very strong one such that any two differing populations must be incomparable to each other. A view similar to this has been defended by Ralf and Bader (2022).[7]

Now, Strong Person-Affecting Views are very strong indeed, and one might balk at the idea that any two differing populations must be incomparable. Of course, almost every view in population axiology has some implausible implications (see (Arrhenius, 2000) who proved how a number of plausible principles cannot by jointly held). So whichever view one chooses to adopt, one better be prepared to bite some bullets. And in this case, we might have good reason to accept even **Strong Person-Affecting Views**, since they are the only set of views, out of all the axiologies surveyed so far, capable of resisting **Beneficent Replacement**.

That being said, it would be nice if our axiology didn't have *quite* so much incomparability as the Strong Person-Affecting Views. Moreover, it would also be nice if our axiology went *further* in saying that replacing humanity entirely with supers is not just *incomparable* to a population of humans and supers, but that a population of only supers is also *worse* than a population of humans and supers.

[7] Bader's view ("Same-Number Person-Affecting Utilitarianism) is similar to the one outlined above because Bader would also say that each of the X_i are incomparable with each other (but only because the populations have different numbers of people). In some sense, Bader's view is weaker than the one outlined above (as long as two populations have the same *number* of people, you can compare them). But in another sense, Bader's view is much stronger. The Strong Person-Affecting Views I have outlined here are consistent with the idea that you can make the world better by adding a person to a population while increasing everyone's well-being, whereas Bader's view implies that such a world would be incomparable to one where we do not add that person.

So on a wide variety of views—Totalism, Averagism, Prioritarianism, Critical-Level Utiliarianism, and Strong and Weak Person-Affecting Views—we are left with no reason to think that the world would be worse off if humanity were replaced with supers. And on many of these views, we have strong reason to think that the world would in fact be better if humanity were replaced. If we want to justify future human flourishing, we need to use a different theory. We will soon see that our theory, like the Strong Person-Affecting Views, will have to tolerate *some* incomparability (but it will not be nearly as ubiquitous), but unlike any of the theories discussed here, our theory will imply that it would be worse overall to replace humanity entirely with supers. Developing and motivating this theory will be the task of the next section.

3 Group prioritarianism

Welfare matters, but how it matters makes all the difference in how we compare populations. In fact, many of the welfarist views touch upon important considerations that point towards the value of future human flourishing. For example, Prioritarians notice that the amount of *good* one's well-being contributes to the overall good *diminishes* as one has more and more well-being. This is a step in the right direction because it implies that benefiting a human (who is usually at a lower level of well-being than a super) by even a bit can matter more than benefiting a super by a lot. But Prioritarians still can't stop **Beneficent Replacement** because the good of bringing a super from non-existence to good existence will always outweigh the good of bringing any other human into the best of human existence.

At this point, one may have been naturally inclined to think a Person-Affecting View would be helpful. Many Person-Affecting Theorists would say that one *does not* make the world better by bringing a super from non-existence to good existence. But at the same time, they would have to say that one would not make the world better by bringing a human from non-existence to good existence. So the Person-Affecting Theorist gives no good reason for thinking that it is better to bring a human into existence *instead* of bringing a super into existence.

I want to bring in the considerations of both the Prioritarian and the Person-Affecting Theorist, but apply just " one weird trick" to resist **Beneficent Replacement**. The weakness of both the Prioritarian and the Person-Affecting views are that they do not take into account *group welfare* as something distinct from the well-being of the individuals of the group. In fact, one might care a lot about group welfare. Consider the following examples:

> **Job offer** Two people, A and B, are equally well off as individuals, and equally qualified for the same job. But A belongs to a historically privileged group and B belongs to a historically marginalized group. Supposing we can benefit either by the same amount by giving them a job offer (and suppose no other *individual* apart from these two would be benefited), one might have the intuition that it would do *more good* to benefit the one from the historically marginalized group. One possible reason for thinking so is that benefiting a member of the worse off group doesn't just benefit that individual, but it benefits

their *group*. And for the same reason one might think it more fair to seek to benefit the worse-off of two individuals (other things being equal), one might also think it would be more fair to seek to benefit the worse-off of two groups (other things being equal).

Save the polar bears The Polar bear population is dwindling and you have two choices. One is to set up a nature preserve to save the polar bears. Doing so would allow 100 polar bears to live with 5 units of well-being per year. The other option is to decimate their habitat to build a new mall. The new mall will benefit 1000 humans (whose total and average well-being far exceeds the polar bears') by 1 unit of well-being per year, but at the expense of all polar bear well-being. Even though building the mall would increase total and average well-being, one might have the intuition that it would be *worse* to benefit humanity at the cost of eliminating the polar bears.

Reparations Two groups of people are now equally well off. However, in the past, one group was oppressed by the other and had significantly less aggregate well-being as a result. One might have the intuition that it would be *better* if some of the well-being in the oppressing group were redistributed to the historically oppressed group, even if total well-being remains constant.

In each of these examples, Totalist, Averagist, and Prioritarian Welfarist theories will say that it doesn't matter what one does in Job Offer and Reparations, and only something like Prioritarianism *might* say something like saving the Polar Bears is better than making the mall. Examples like these, then, are usually set up as counter-examples to welfarist axiologies. One might point to these examples to suggest that there is something *other* than well-being that matters such that the non-welfare maximizing option may be the better one. For example, perhaps there is intrinsic value to *diversity* (as in **Job Offer**), or *biodiversity* (as in **Save the Polar Bears**), or in *justice* (as in **Reparations**).

However, although one can explain these intuitions by appealing to values as disparate as biodiversity and justice, the welfarist need not do so. In fact, we can adopt a more unified explanation that appeals only to the existence of group welfare, and the need to be fair to groups. In other words, we can model the value of things like *diversity*, *bio-diversity*, and to some extent, *collective justice*, within a welfarist axiology. The basic idea is that we can increase *total goodness* at a faster rate when welfare is distributed to members of less well off *groups*.

Let us call the view just described *Group Prioritarianism*. Here is one formal statement of the view that will apply for our purposes. To begin, let X be a population, where a "population" is represented as a vector $< w_s(x_s), w_h(x_h) >$, where x_i is the percentage of our available resources given to group i (and where x_s is the percentage of resources given to the supers and x_h the percentage of resources given to humans), w_i is a function from x_i to the aggregate well-being of group i. For now, we can simply assume that this aggregate well-being function is Totalist for ease of

exposition.[8] Let O be our function on a population, X, to its overall goodness. Here's our definition of *Group Prioritarianism*

$$O(X) = g(w_s(x_s)) + g(w_h(x_h))$$

where g is our concave "goodness" function that tells us how much good the well-being of a group contributes to the overall good. The fact that our goodness function, g, is concave means that the greater a group's aggregate well-being, the less that group's marginal well-being would contribute to overall goodness. Supposing that these supers are able to convert resources into well-being 1000 times more efficiently than a human (i.e. $w_s(x) = 1000 * w_h(x)$), then if there are already many supers enjoying extraordinary levels of well-being such that their contribution to the overall good is high, then it would do *more good* to give the same resource to a human even if the total amount of *well-being* would be less than if it had been given to a super.

Now, if our function g is concave enough, then we can see that there is a point in which a beneficent replacement will make things worse, and not better. This will happen when there are so many supers already enjoying so much aggregate well-being that an additional 10 supers will not matter as much as keeping one additional happy human. In that case, **Beneficent replacement** would be false.

For example, suppose our function g is just the square root function. In this case, we can calculate the point at which things will get worse when we transfer human resources to AI resources by just solving the following constrained optimization problem. Let x be the percentage of the universe's resources that go to humanity and y the percentage of the universe's resources that go to the supers. The important thing is that an AI can produce 1000 times more well-being for each resource to support a human, so we will just set $w_h(x) = x$ and note that $w_s(x) = 1000 \times w_h(x)$.[9] Now the question is to optimize $z = \sqrt{w_s(y)} + \sqrt{w_h(x)} = \sqrt{1000 \times w_h(y)} + \sqrt{w_h(x)} = \sqrt{1000 \times y} + \sqrt{x}$, with the constraint that the percentage of resources that go to the supers and humans add up to 100 (i.e. $100 = x + y$).[10] The answer is that $x \approx 0.1\%$ and $y \approx 99.9\%$. So on **Group prioritarianism** with a square root function, we see that the ideal distribution is to give humanity just a tiny slice of the universal pie.[11] Strikingly, if we had instead stipulated that the supers can consume resources *ten thousand* times more efficiently, then the optimum distribution would be 99.99% to the supers and 0.01% to humanity, just as Bostrom and Shulman (2020) guesstimated.

At present, **Group prioritarianism** is just a *sketch* of a family of views that all differ in terms of the concave function g that they adopt, and what *groups* they

[8] A Prioritarian function could also work, but nothing much turns on whether our aggregate well-being function is Totalist or Prioritarian. In Sect. 4.3, we will consider some reasons to rethink our aggregate well-being function and consider another one.

[9] This is because we are assuming that a resource that goes to a human can be used to support 10 AI with 100 times more well-being each.

[10] Here, x is the percentage of resources that goes to humanity and y is the percentage of resources that goes to the supers.

[11] The calculations were made by a simple python script.

admit. First, the concavity of our function g matters. We just happened to use the square root function because of its simplicity, but we could have also easily used the even more concave natural log function. In that case, the optimal distribution would turn out to be close to 50/50. Of particular importance, however, is how we individuate groups. If we adopt the view that "everyone is a special individual" that belongs to their own group, then **Group prioritarianism** just becomes Prioritarianism. If, on the other hand, we adopt the view that "we are all one big happy family" in one group, then **Group prioritarianism** just devolves into something like total Utilitarianism or some other standard welfarist axiology (depending on how we aggregate the well-being of a group). And as we have seen above, all standard welfarist axiologies favor a population exclusively of supers, and so it really matters how we individuate our groups.

So group individuation matters, but for precisely that reason, group individuation is not arbitrary. For example, just because we can gerrymander human populations based on the number of hairs on one's head, or based on some other normatively irrelevant feature, it doesn't mean that any individuation of groups all have equal say on how we ought to redistribute resources. It is more plausible that we should give more resources to groups that can be individuated based on patterns of historic injustice (for example) as opposed to distributing resources to ensure that people whose last name starts with "k" get as much as the rest. Though it isn't easy to find the precise theory for how to individuate groups (and it is not the place of this paper to develop such a theory), one seemingly plausible heuristic for any such theory is that the normatively relevant features that individuate groups should be features that help explain vast discrepancies in well-being between groups. For example, if well-being differs drastically along gendered or racial lines, and the different groups experience differing levels of well-being *because* they are in one group as opposed to another, then it seems that individuating groups based on racial and gendered lines is normatively relevant. In our case, we are considering populations which have at least two groups with *vastly* different levels of well-being, and the best explanation for why these two groups have different levels of well-being is that one group consists of supers (capable of super-human well-being) and another group does not. So if any group distinctions are normatively relevant, then the distinction between humans and supers is normatively relevant.[12]

By now, it should be obvious how **Group prioritarianism** draws inspiration from Prioritarianism. However, I also mentioned at the beginning of this section

[12] Although this heuristic alone is enough to see why humans and supers belong to different groups, it is worth emphasizing that this heuristic is not a *necessary* condition for individuating groups. For example, we often take a pair of minority groups to be distinct even if the welfare levels *between those two groups* are similar. Similarly, one might think it important to distinguish a minority group from the majority group when the two groups have comparable well-being (for example, we may distinguish between the Cantonese speaking minority in China from the non-Cantonese speaking majority, and recognize that some priority should be given to accommodate Cantonese speakers (e.g. through Cantonese speaking schools, news channels, etc). Similarly, one might think that if we are surrounded by a large population of AI with only *human-level* well-being, then although humans and AI have comparable levels of well-being, the two groups ought to be counted as distinct, and special priority should be given to accomodate the human minority (e.g. through human schools, human recreation centers, human-hospitable planets, etc.)

that we also want to draw insights from the Person-Affecting Views. Just as Person-Affecting Views are characterized by the slogan “we are in favor of making people happy, but neutral about making happy people", so too do we want **Group prioritarianism** to advocate the slogan “we are in favor of making groups happy, but neutral about making happy groups".

Let us call the principle that populations with different groups are *incomparable* the “Strong Group-Affecting Principle". The idea, then, is that one cannot simply make the world better by creating a new group. This is especially implausible if one does nothing to increase the welfare distribution among people, but only change how people are grouped together. For example, if one can compare populations that differ in what groups they contain, then it should be possible to increase overall goodness by splitting a monolithic group, G, into two groups G_b and $G_{\neg b}$ containing bald people and non-bald people respectively by systematically oppressing people who are bald by just a little bit. In that case, we can imagine that the total goodness of that population will go from $g(1000)$ to $g(500 - \epsilon)$ for the bald group and $g(500)$ for the non-bald group. Now, recall that for any increasing concave function g, $g(x) + g(y) \geq g(x + y)$ (for positive x and y). If ϵ is small enough and g is concave enough, then the resulting split population would contain more goodness than the monolithic population. But this is ridiculous. One cannot make the world better by arbitrarily forming a group via oppression. For reasons like this, we must adopt a Group-Affecting Principle.

Granted, this principle is very strong, and it has some very unsavory consequences. For example, it would imply that, although one would not make the world *better* by forming another group via oppression, it would not make the world worse either.

At this point, one might consider adopting a “Weak Group Affecting Principle", which posit group affecting considerations as only one (albeit an important one) of many normative dimensions. So on this view, forming a group via oppression doesn’t make the world any better in the group-affecting dimension because there is no single group that becomes better off, and forming the group via oppression makes things *worse* in other dimensions (like reducing total and average welfare). Thus, it is compatible with the “Weak Group Affecting Principle" to say that the world is *worse* overall if one forms a new group via oppression.

However, in our context, adopting a “Weak Group Affecting Principle" has problems in its own rights. For one, adopting a Weak Group-Affecting Principle makes it unclear whether it really is bad to replace all humanity with supers. Of course, it will be very bad on the important Group-Affecting dimension because it severely harms the human group. On the other hand, the Group-Affecting dimension is just one normative dimension, and it is unclear whether the world may just be better overall because the world is much better off in some other compensating dimension (e.g. the dimension of total and average well-being).

So we have a dilemma. Populations with different groups are either comparable or incomparable. If they are comparable, then it is unclear whether adding a new group of supers makes things *worse* overall. Indeed, if we accept a Weak Group Affecting Principle, adding a new group of supers that completely replaces humanity may still end up making the world *better*.

However, if worlds with different groups are *incomparable*, then we get weird results like saying that forming a new group via oppression does not make things *worse*. Moreover, it would also imply that a world with only humans is incomparable to a world where there are humans and supers.[13]

In the interests of humanity, I am inclined to adopt the Strong Group-Affecting Principle in order to deny that we have any reason for creating a new group of supers, and in order to preclude the possibility that, having created the new group of supers, we may still need to give all our resources to the supers at the expense of humanity.[14]

But now, having accepted the Strong Group-Affecting Principle, one might wonder whether we can just rerun the argument from **Beneficent replacement** by just stipulating that each human is replaced by 10 supers that belong to a *new group*. Our Strong Group-Affecting Principle would say that each replacement will not make the world any *better*, but it also would say the world won't be any worse. So in this way, our Strong Group-Affecting Principle stands to this revised version of the Beneficent Replacement argument in the same way the Strong Person-Affecting Principle stands to the original version of the argument. The principle succeeds in resisting the argument to the point that it helps us deny **Beneficent replacement**, but it doesn't go so far as to say that the world is worse if humanity is replaced in this new way.

However, unlike the situation where we can choose to create 10 new supers, it is not really in our power to make each new super part of a new group. To do so, we would have to treat each group in a completely different way socially. So this revised version of the Beneficent Replacement argument is, in an important sense, moot.[15]

Hopefully by now, one has a good sense of **Group prioritarianism**—how it is motivated, and how it can resist **Beneficent replacement**. The next section discusses some objections to the view.

[13] Note that the Strong Group Affecting Principle would also imply that a world of *only* supers is incomparable with a world of *only* humans. However, although a future containing only supers is something we can choose to create, we cannot create a *world* with only supers. This is because, in the actual world, the group of humanity already exists, and so now we can only choose between three kinds of worlds (1). a world with only humans, (2). a world where humans and supers exist at the same time, and (3). a world where the group of humans goes extinct and are replaced by a group of supers. In all three worlds, the human group exists. The Strong Group Affecting Principle tells us that (1) is incomparable to (2) and (3), but it will tell us that (2) and (3) are still comparable with each other. As an analogy, one may consider a (Very Strong) Person Affecting Principle that tells us that any worlds with different people are incomparable. Now consider three kinds of worlds—(1). Only Alice exists, (2). Alice and Bob exist together, (3). Alice exists and is killed to produce Bob. Our (Very Strong) Person Affecting Principle would tell us that (1) is incomparable with (2) and (3), but (2) and (3) are still comparable to each other.

[14] It should be noted that, although both the Strong Person-Affecting Principle and the Strong Group-Affecting Principle would say the world is not *worse* when we add in a new group of supers, only the Strong Group-Affecting Principle would say that, once the group of supers is added, the world would be *worse* if humanity were completely extinct.

[15] Thanks to the editor for raising this point.

4 Objections to the view

4.1 Objection 1: How do we individuate groups?

The viability of this approach depends in large part on how we ought to individuate groups. In the last section, I mentioned that the normatively relevant features that individuate groups should be features that help explain vast discrepancies in well-being between groups. One worry is that this principle can be abused to justify gerry-mandered groups. For example, one might wonder if the group containing people with *exactly* the same amount of well-being as Bob is a normatively relevant group. Perhaps one can argue that one can explain why anyone in this group has a different level of well-being from any other group because people in that group all share the feature of having the same amount of well-being as Bob and no one else. If we can individuate groups in this way, and if no two people have *exactly* the same amount of well-being, then **Group prioritarianism** threatens to just reduce into ordinary Prioritarianism.

To be clear, this objection is not an objection against **Group prioritarianism** *per se*. Rather, this is an objection to a hypothetical theory for group individuation. *If* we had a theory that said that groups ought to be individuated by virtue of their differences in welfare (no matter how small), then the gerry-mandered group above would count as a counter-example to the theory—it would be an example of a group that satisfies the principles of that theory without being an example of a normatively relevant group. If that's the case, then we need a new theory of group individuation, not a new axiology that doesn't rely on group individuation.

So what is our theory of group individuation, and what makes a group normatively relevant? Again, it is beyond the scope of this paper to give a fully worked out theory; however, we can say a few instructive things about how we would go about constructing such a theory. The most important point is that our theory of what makes a group "normatively relevant" cannot be constructed apart from our judgments about which allocation of resources across groups is better than another. If we did so, our theory would almost assuredly go wrong. To explain this methodological point, it would be fruitful to compare our approach to the better known approach David Lewis takes in his analysis for counterfactuals.[16]

For Lewis, a counterfactual "if *A* were the case, then B would be the case" is true just in case B is true in all the closest worlds where A is true, where a world *w* is *closer* to actuality than *w′* iff *w* is more similar to actuality than *w′*. And how does one know whether one world is more similar than another world? We do not do so by referring to a theory of "comparative similarity" that makes no reference

[16] The following discussion is inspired by a similar discussion found in Williamson (2009) which compares the notion of "safety" in Williamson's theory of knowledge with the notion of "comparative similarity" in Lewis's theory of counterfactuals.

to counterfactuals themselves. All such theories are bound to get things wrong.[17] Rather, our judgments on whether any two worlds are more similar to actuality should be determined by which counterfactuals we think are true.

Furthermore, even before Lewis gives his theory of comparative similarity, his "fully general" theory of counterfactuals is not devoid of content (Lewis, 1986). For example, the fully general theory can be used to elucidate interesting things about the *logic* of counterfactuals.

Similarly, **Group prioritarianism** is an axiology that essentially relies on the notion of a *normatively relevant group* in the same way Lewis's theory of counterfactuals relies on the notion of *comparative similarity between worlds*. Just as one cannot determine *comparative similarity* without reference to our intuitions about counterfactuals themselves, one cannot determine which groups are *normatively relevant* without reference to our intuitions about axiology either. For example, in determining whether "people who have the exact level of welfare as Bob" counts as a normatively relevant group, we must refer to our judgments about group redistribution. For example, we must ask whether it's true that transferring welfare from people who belong to another group to the people who belong to Bob's group makes things better overall just because they belong to Bob's group. The answer to that question is "obviously not", for the same reason that a person who claims that they should be given priority for being in "a minority of one" can only be arguing in bad faith. So for that reason, these gerry-mandered groups are ruled out as normatively relevant.

And just as how Lewis's general theory for counterfactuals can give us useful insights about the logic of counterfactuals even without giving us an analysis of "comparative similarity", so too does our group-based axiology give us useful insights about the logic of group redistributions. For example, **Group prioritarianism** tells us that one can better maximize overall welfare by giving priority to groups which consume resources less efficiently. And it tells us that, given that the function g is a log function, and given that the welfare levels between humans and supers stipulated above, one ought to split our total resources 50–50 between these two groups. And most crucially, it tells us that **Beneficent replacement** is false.

4.2 Objection 2: Egyptology?

In Parfit's *Reasons and Persons*, he gives an influential objection against the Averageist view (Parfit, 1984, p. 420). The objection is that, if Averageism is true,

[17] For example, some attempt to cash out "comparative similarity" between worlds in terms of overall space-time likeness, but such an analysis, combined with Lewis's theory of counterfactuals, would imply that the counterfactual "If the president pressed the nuclear launch button during the Korean War, then the button would have malfunctioned" is true simply because a world where a small malfunction happens is overall more like our world than one where there had been a nuclear bomb detonated during the war. Such an attempt to analyze "comparative similarity" without reference to counterfactuals themselves is misguided. Rather, one should take from this example that the falsity of the above counterfactual tells us that a world where there is a malfunction is *not* more similar to actuality than a world where there isn't.

then the continuing of the human race, with the certainty that the future will contain many many flourishing individuals, would be considered impermissible if we knew that the very pinnacle of human well-being happened already in Ancient Egypt. The idea is that, even if future generations will have good lives, the addition of future flourishing humans can only bring down the average because nothing can match the heights of Ancient Egyptian welfare. Thus, on the Averageist view, it would be impermissible to continue the human race. Moreover, in deciding whether or not we should continue the human race, one must first do some Egyptology to find out exactly how well off Ancient Egyptians really were so as to ensure that we do not lower average human well-being. Of course, this result is ridiculous on two accounts. First, it is implausible that one can make the world *worse* by bringing into existence flourishing human lives. Secondly, it is implausible that in deciding whether we *should* continue the human race, we should do some Egyptology. What happened in Egypt stays in Egypt, and it should have no effect on the goodness or badness of continuing human flourishing.

Similarly, our decisions to give certain goods to different people will also be dependent on historical facts on the overall well-being of particular groups. For example, suppose (for simplicity) that the world contained only two groups of people—Egyptians and non-Egyptians. Suppose Asim and Brandon are equally well off and that Asim is an Egyptian and Brandon a non-Egyptian. If **Group prioritarianism** is true, then the relative goodness of giving Asim a chocolate bar over giving giving Brandon a chocolate bar depends on the aggregate well-being of their respective groups. If, for example, the Ancient Egyptians had a golden age and their aggregate well-being exceeded that of Brandon's group, then **Group prioritarianism** tells us that we should give the chocolate bar to Brandon. But this is just to say that, in order to know whom to give a chocolate bar, we should do some Egyptology.

This result, however, does not strike me as bad as Parfit's Egyptology objection to Averageism for two reasons. The first is that this objection does not tell us that giving benefits to *either* group will make things *worse*, and it certainly doesn't tell us that prolonged existence of any group will make things *worse*. Secondly, although this example implies that Egyptology is necessary to know which action would produce the most good, I think this sensitivity to historical facts is a feature and not a bug.

In fact, one weakness of standard welfarist axiologies is that they are insensitive to historical facts about group welfare. For example, Prioritarian axiologies that seek to benefit the worst off may cease to recommend that reparations be paid to certain historically disadvantaged and oppressed groups so long as the *current* welfare of the individuals of that group matches those of their former oppressors. **Group prioritarianism**, on the other hand, would say that any additional goodness given to a historically oppressed group doesn't disappear the moment the group ceases to be oppressed. After all, reparations are motivated to compensate for past injustices, not current injustices. So, in deciding how to distribute resources, learning a bit of history, and perhaps some Egyptology, can help.

4.3 Objection 3: Implausible verdicts?

Depending on how we fill out the details of **Group prioritarianism**, the view implies some wild conclusions that may undercut its initial motivations. Of course, some counter-intuitive conclusions are to be expected—for example, that there will be cases where benefiting a better off individual by a small amount can be better than benefiting a worse-off individual by a large amount, provided that the better off individual is in a much more disadvantaged group. But such a conclusion is just something we have to accept if we want to avoid situations where human needs are completely crowded out by needy supers who happen to have just below average human well-being. However, the following example is much more worrying for proponents of **Group prioritarianism**:

> **Neglect the masses**: Imagine there are two groups: The Few and The Masses. The Few consists of a handful of fabulously well-off royalty. The Masses consist in an astronomically large population of people whose lives are just barely worth living. You now have a choice on how to spend a rare resource. The resource can either double the well-being of everyone in The Masses, or it can bring into existence a new member of The Few whose life is just okay. Which action would do the most good?

Up till now, we have been assuming that aggregate group welfare is determined by some kind of additive axiology like Totalism or Prioritarianism. Both those views, however, famously imply the Repugnant Conclusion, which states that a population of people whose lives are barely worth living is better than a population of people whose lives are great, provided that the first population is much larger than the second. So, assuming that aggregate group welfare is determined by either Totalism or Prioritarianism, we can construct a case where The Masses is a better population than The Few. But in that case, if g is concave enough and the population for The Masses is big enough, **Group prioritarianism** implies that we should benefit The Few by bringing into existence an okay life instead of doubling everyone's well-being among The Masses. This result seems antithetical to our initial motivation to construct an axiology that benefits disadvantaged groups. So although our view can capture many of our intuitions regarding the value of group welfare, our view will not be able to accommodate them all.

There are two ways to respond to this worry. One way is to find a better way of aggregating group welfare (a different function for w). For example, perhaps aggregate group welfare should be determined by the "Variable Value" view defended by Hurka (1982). On that view, when the population is low, aggregate value is best increased by increasing the total. But as the population grows, adding more people adds far less value, in which case it would be better to increase the average. Such a view would resist the Repugnant Conclusion and say that there is a point where a population, X, is happy enough and large enough such that there is no other population, Y, with very low average well-being that is better than X by virtue of having a much greater population size.

Indeed, the Variable Value view may find its most natural home in the context of playing the role of the group aggregate function in **Group prioritarianism**. Hurka

(1982) even begins his paper by hinting at the value of having diversity of groups by quoting Thomas Aquinas: “Just because an angel is better than a stone, it does not follow that two angels is better than an angel and a stone" (Aquinas, 1975, III, 71). However, the Variable Value view alone cannot resist **Beneficent replacement** unless it takes group individuation as being normatively significant, and this is something left out of Hurka’s picture. Indeed, if supers and humans are treated as one population, the Variable Value view, by itself, would imply that once we have enough happy supers, we should not even allow for any additional humans to exist, even for free, simply because adding more people brings very little value and lowering the average at this point brings a lot of disvalue. But if we take the Variable Value view as playing the role of being the aggregate group welfare function, we avoid this result.

But Variable Value views have their own problems, and indeed, any view that rejects the Repugnant Conclusion has to pick from a wide array of unattractive features (Arrhenius, 2000). So another way to respond to our problem is to embrace the result that we should benefit The Few over The Masses. After all, from the perspective of insects, they are the Masses and humanity is The Few. If we concoct a view that systematically benefits The Masses over The Few, then it may be the case that all human resources should go to benefiting insects instead (or perhaps to a swarming population of AI that have even less well-being than an insect.).

Then again, any view that *always* benefits The Few over The Masses would be to our disadvantage once the first supers come into existence. For from their perspective, they are The Few and we are The Masses. Thus, humanity is in an awkward position. Views that seek to increase the population of groups with the highest average well-being would result in humanity being crushed by the overwhelming demands from above (i.e. we should give everything to supers). And views that seek to do everything to increase the welfare of groups with the lowest average well being would result in humanity being crushed by the overwhelming demands from below (i.e. we should give everything to insects). I think the view sketched in this paper is a reasonable middle ground. The view may at times give us unintuitive consequences—but sacrifices must be made for the sake of human survival.

5 Conclusion

If super-beneficiaries ever come, the question about whether the human species should be replaced will be the most important question humanity will have to answer. It is no help that all welfarist axiologies until now advocate for human replacement. For this reason, defenders of humanity may think that there are more goods in this world than just welfare, and that diversity is one of them. In this paper, I accommodate this intuition that diversity is valuable without positing diversity as a value. In the axiology developed here, distributions of resources over diverse groups are valuable because *well-being* is more valuable when given to less well-off groups.

In many ways, however, the view developed in this paper is under-specified. What is presented here is instead a general axiological framework that is capable of resisting **Beneficent replacement**. But in order for this axiology to give more concrete

guidance on how to distribute resources, we need to specify (1) what groups there are, and (2) how to aggregate group well-being. These are not easy questions to answer. To answer (1), we need to do more metaphysics and sociology. To answer (2), we need to do even more population ethics. Much work needs to be done to fully flesh out **Group prioritarianism**, but it is my hope that the framework presented here may be able to absorb the insights from many disparate areas in the humanities to help settle the question of how to best distribute our resources. But whatever the answers to those two questions may be, I'm confident that one recommendation of our framework will still stand: when the supers come, humanity should not be replaced.

Acknowledgement This paper was largely written during my time as a research fellow at the Center for AI Safety. Many thanks to Mitch Barrington, Cameron Domenico Kirk-Giannini, William D'alessandro, Simon Goldstein, Jacqueline Harding, Nick Laskowski, Harry Lloyd, Robert Long, Nate Sharadin, and Elliot Thornley for their fantastic input at the Center. Thanks also to John Hawthorne and Jeff Russell for great discussion. Special thanks are due to Ben Levinstein who provided many comments on several early drafts.

References

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics and Philosophy, 16*(2), 247–266. https://doi.org/10.1017/s0266267100000249

Cohen, G. A. (2012). *Finding oneself in the other*. University Press.

Fischer, B., & Sebo, J. (2024). Intersubstrate welfare comparisons: Important, difficult, and potentially tractable. *Utilitas, 36*(1), 50–63. https://doi.org/10.1017/s0953820823000286

Goldstein, S., & Kirk-Giannini, C. D. (2023). AI wellbeing.

Hurka, T. (1982). Value and population size. *Ethics, 93*(3), 496–507. https://doi.org/10.1086/292462

Lewis, D. (1986). *Philosophical papers* (Vol. Ii). Oxford University Press.

Mogensen, A. (2023). Welfare and felt duration. *Global Priorities Institute Working Paper Series*.

Narveson, J. (1973). Moral problems of population. *The Monist, 57*(1), 62–86. https://doi.org/10.5840/monist197357134

Nebel, J. M. (2021). Conservatisms about the valuable. *Australasian Journal of Philosophy, 100*(1), 180–194. https://doi.org/10.1080/00048402.2020.1861037

Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.

Parfit, D. (1984). *Reasons and persons*. Oxford University Press.

Ralf, M., & Bader. (2022). Person-affecting utilitarianism. In G. Arrhenius, K. Bykvist, T. Campbell, & E. Finneron-Burns (Eds.), *The Oxford Handbook of Population Ethics.* Oxford University Press.

Ross, J. (2015). Rethinking the person-affecting principle. *Journal of Moral Philosophy, 12*(4), 428–461. https://doi.org/10.1163/17455243-01204004

Samuel, S. (2018). *Why worry about future generations?* Oxford University Press.

Shulman, C., & Bostrom, N. (2021). Sharing the world with digital minds. Rethinking moral status, pp. 306-326.

Thomas, A. (1975). *Summa contra gentiles*. University of Notre Dame Press.

Timothy, W. (2009). Probability and danger. *The Amherst Lecture in Philosophy, 4*, 1–35.