The Utility Monster

BLINDED FOR REVIEW

This paper develops a new impossibility result that connects population axiologies to welfare axiologies. The result is centered around a puzzle about the *utility monster*, defined as an individual whose lifetime quantity of welfare goods exceeds the overall quantity of welfare goods across an entire population of excellent human lives. The puzzle is about how to avoid MONSTROSITY, the claim that for any world containing only excellent human lives, there's a better world containing only a lone utility monster. I show how the puzzle can be formulated with minimal assumptions about which theory of welfare is correct, how welfare and value are structured, and even whether the utility monster has an especially good life. The argument demonstrates that that it's impossible to deny MONSTROSITY while accepting several other plausible premises. The upshot is that any future axiology must—in one way or another—be monstrous.

Introduction

This paper develops a puzzle that connects population axiologies to welfare axiologies. The puzzle is centered around the following claim:

> MONSTROSITY
> For any world A containing only excellent human lives, there's a better world Ω containing only a utility monster.

A *utility monster*, in the sense at play here, is an individual whose lifetime quantity of welfare goods exceeds the overall quantity of welfare goods across the entire population of A. You might initially be inclined to dismiss the possibility of utility monsters. But note that the definition above appeals only to the quantity of welfare goods accrued over a lifetime,

leaving open whether that massive amount of welfare goods converts to a monstrously high welfare level.[1]

The puzzle is about how to develop a satisfactory axiology that avoids MONSTROSITY. This is analogous to the puzzle generated by the most famous (or infamous) claim in population ethics:

> REPUGNANCE
> For any world A containing only excellent lives, there's a better world Z containing a much larger population consisting only of lives barely worth living.[2]

A great deal of population ethics has been focused on finding ways of resisting REPUGNANCE. But—as I'll explain—nearly all prominent strategies for resisting REPUGNANCE permit MONSTROSITY. Therefore, those who wish to resist MONSTROSITY will need to find new strategies beyond those used to resist REPUGNANCE alone. Furthermore, solving the puzzle of MONSTROSITY requires thinking about not only the philosophical questions traditionally addressed within population ethics, but also analogous questions about the structure of welfare and the values of individual lives.

The centerpiece of this paper is a new impossibility result for population ethics and theories of welfare. I'll explain how the negation of MONSTROSITY is incompatible with accepting several other premises, each of which is plausible and intuitive. Furthermore, I'll explain how the impossibility result rests on only minimal assumptions: it leaves open questions about how to weigh different kinds of welfare goods against each other, the

[1] The term 'utility monster' originates from Nozick [1974], though Nozick defined the utility monster as a creature that's extremely efficient at converting resources into welfare, and invoked the utility monster to motivate the normative significance of equality and rights. By contrast, this paper defines 'utility monster' in a way that leaves open questions about whether the utility monster has an especially high welfare level and whether equality and rights are normatively significant.

[2] See Parfit [1984] for the original discussion of REPUGNANCE.

structure of welfare and value, whether or not social goods matter, and whether or not lonely utility monsters can have excellent lives.

Here's the plan. **§1** defines and defends the possibility of the utility monster; **§2** presents the impossibility result; **§3** discusses responses to the impossibility result.

§1 The Utility Monster

The idea of a utility monster may initially strike some as far-fetched. But a *utility monster*, in the sense at play here, is simply an individual who accrues a *monstrous quantity* of welfare goods, meaning its lifetime quantity of welfare goods exceeds the total quantity of welfare goods accrued by all the individuals in world A (a population of excellent lives). In what follows, I'll explain this idea in more detail.

Utility, Welfare, Value

For the purposes of this paper, it will be useful to distinguish what I'll call 'utility', 'welfare', and 'value'. I'll use 'value' and 'welfare' in the standard ways, but I'll use 'utility' in a non-standard way:[3]

utility	quantity of welfare goods/bads
welfare	goodness/badness of a life (for an individual)
value	goodness/badness of a world

In other words, 'value' concerns impersonal value/good-simpliciter, 'welfare' concerns personal value/prudential value/good-for, and 'utility'—as I'm using the term—concerns the quantity of things (pleasure, satisfied-desires, knowledge, etc.) that contribute to welfare. Consider, for

[3] There's no single standardized meaning for the term 'utility'. In many contexts, 'utility' is used synonymously with 'welfare'. In economics, 'utility' is often used to denote the numerical representation of an individual's preferences over a set of alternatives. In formal ethics, 'utility' is often used to denote the numerical representation of welfare levels.

illustration, an individual who undergoes a series of pleasurable experiences. Here, 'value' denotes how much that pleasure contributes to the goodness of the world, 'welfare' denotes how good the life with that pleasure is for the individual who lives it, and 'utility' denotes the overall quantity of pleasure.[4]

I use 'utility' in this way because there's no canonical term for expressing quantity of welfare goods. This may be because it's often taken for granted that welfare is simply a linear function of quantity of welfare goods. The implicit assumption in the welfare literature has been "welfare totalism," according to which amount of welfare equals amount of utility (just as totalism, in population ethics, holds that amount of value equals amount of welfare). A central aim of this paper is to show how this assumption leads to the puzzle of MONSTROSITY.

When we distinguish value, welfare, and utility, it becomes clear that there are substantive questions about how they relate to each other. This is obvious when thinking about value and welfare, since different population axiologies make different claims about the relationship between these quantities. But analogous questions also arise when thinking about welfare and utility. In fact, as I'll discuss later, a natural way of resisting MONSTROSITY is to hold that there's an asymptotic function from utility to welfare (meaning that there's an upper-bound on welfare levels, even if there's no upper-bound on utility).

Once we define the utility monster in terms of utility (as opposed to welfare or value), it's hard to deny that utility monsters are possible.[5] It's

[4] For simplicity, I'll frame the whole discussion in terms of 'welfare goods' (rather than 'welfare goods/bads'). Whenever I ascribe utility / welfare / value to individuals, I'll mean lifetime utility / welfare / value. I'll express utility in terms of 'gains' and 'losses' (to an individual), welfare in terms of 'benefits' and 'harms' (to an individual), and value in terms of what's 'better' or 'worse' (without relativization to an individual).

[5] Parfit [1984: 389] says that the utility monster is "deeply impossible." But he has in mind a welfare monster that's unimaginably efficient at converting resources into welfare. That skeptical sentiment doesn't obviously generalize to the sort of utility monster at play here.

plausible that individuals can live arbitrarily long.[6] And with an arbitrarily long lifespan, an individual could accrue greater and greater amounts of pleasure, knowledge, satisfied desires, and nearly any other goods. Even if there's an upper-bound on the amount of welfare goods that can be accrued at a single moment, there's no obvious reason for thinking that there's an upper-bound on the quantity of welfare goods that can be accrued over an arbitrarily long lifetime. And even if there's an upper-bound on possible welfare levels, there's no obvious upper-bound on the lifetime quantity of welfare goods that an individual can accrue.

Utility monsters occasionally make appearances in the philosophical literature. But in almost all cases, the focus is really on *welfare monsters*, or individuals with monstrous welfare levels. The most famous example is Nozick [1974], who coined the term 'utility monster'. Other notable examples include Ng [1989], Carter [2011], Pivato [2020], and Stroppa [2024][7] Since you could question whether it's possible for an individual to have a monstrous welfare level, you could question the possibility of the welfare monsters invoked by those authors. But, as I'll discuss more later, it's harder to resist the kind of utility monster defined in this paper.

Although the utility monster is a philosophical thought-experiment, there's a live possibility that some future AI system will, in the relevant sense, be a utility monster. That is, we could imagine an AI that satisfies the requirements for welfare subjecthood and that's able to accrue monstrous quantities of welfare goods. Imagine, for example, that the AI is conscious, has many intensely pleasurable experiences, has many strong desires that are easily satisfied, and acquires vast amounts of knowledge of important

[6] For more discussion, see the "Nomic Limits" and "Personal Identity" responses in §3.

[7] Ng [1989] appeals to a lonely utility monster world to question the intuition that world Z (from REPUGNANCE) is worse than world A. Carter [2011] considers a world with "one mindbogglingly happy, worthwhile life." Pivato [2020] argues for a "no utility monsters" principle: for any population size n, there's some larger world that's better than any n-sized world. Stroppa [2024] uses the utility monster to argue for a version of prioritarianism that places an upper-bound on the amount of value contributed by a single life.

subject-matters. Maybe it's unlikely that such AI systems will be instantiated anytime in the near future, but the possibility makes the idea of a utility monster more concrete.

Population Axiologies

In population ethics, there's little discussion of upper-bounds on welfare levels. Oftentimes, it's tacitly assumed that welfare is unbounded, meaning there's no limit to how high an individual's lifetime welfare level could be. Although my impossibility result will leave open whether or not welfare is unbounded, it's worth seeing how taking welfare to be unbounded quickly leads the most prominent population axiologies to MONSTROSITY.

Suppose welfare is unbounded. Then it follows that for any world A containing only excellent human lives, there's a world Ω containing only a lonely utility monster such that (***a***) Ω has greater total welfare than A, (***b***) Ω has greater average welfare than A, (***c***) Ω has at least as much equality as A, (***d***) the utility monster has a monstrously good life, and (***e***) the utility monster has access to some "higher goods" (such as intellectual and aesthetic appreciation). These consequences mean that the most prominent population axiologies permit MONSTROSITY.[8] Here's a brief explanation.

Let '$W_1 > W_2$' mean 'world W_1 is better than world W_2'. Because of (*a*), **totalism** (where $W_1 > W_2$ iff the total welfare in W_1 is greater than the total welfare in W_2) entails that Ω is better than A. Because of (*b*), **averagism** (where $W_1 > W_2$ iff the average welfare in W_1 is greater than the average welfare in W_2) entails that Ω is better than A. Because of the combination of (*a*)

[8] See Greaves [2017] and Arrhenius & Andersson [2021] for some recent overviews of population axiologies. See Stroppa [2024] for more extensive arguments that these population axiologies entail MONSTROSITY (if welfare is unbounded). Alongside the axiologies listed here, two other prominent responses to REPUGNANCE are (*a*) rejecting the transitivity of 'better-than', and (*b*) adopting a person-affecting restriction on the betterness of worlds. However, neither of those responses provide ways of resisting MONSTROSITY, since (*a*) the impossibility result involves only two worlds, so there's no room for a failure of transitivity, and (*b*) the two worlds in the impossibility result could consist of the same individuals.

and (*b*), **variable value views**[9] (which postulate a concave function from total welfare and population size to value, and which occupy a middle ground between totalism and averagism) entail that Ω is better than A. Because of (*d*), **critical level views**[10] (where $W_1 > W_2$ iff the sum of the differences between each person's welfare and the critical level is greater in W_1 than in W_2) entail that Ω is better than A. Because of (*e*), **lexical threshold views**[11] (where $W_1 > W_2$ iff W_1 contains sufficiently more higher goods and lower goods than W_2, and where some amounts of the former can outweigh any amount of the latter) entail that Ω is better than A.

To develop an axiology that avoids MONSTROSITY, one must appeal to new moves beyond those used to resist REPUGNANCE. A key step, for any such view, is to impose an upper-bound on the amount of value contributed by an individual's welfare. But it's no easy feat to develop a plausible axiology that does this, for any such view will also have to reject one of the premises that figure into the impossibility result.

§2 The Impossibility Result

The following five claims are incompatible (I'll soon briefly explain the premises, and I'll discuss strategies for rejecting them in §3).[12]

1. ANTI-MONSTROSITY

   For some worlds containing only excellent human lives, there's no better world containing only a utility monster.

[9] See Hurka [1983], Ng [1989], and Sider [1991].

[10] See Blackorby *et al* [1997] and Broome [2004].

[11] See Arrhenius & Rabinowicz [2015], Thomas [2018], and Nebel [2022].

[12] Alongside premises 1–5, I'll also take for granted the following background assumptions: (1) there are many levels for utility and welfare (enough to make sense of the difference between very small and very large differences), and (2) for ordinary humans, very small differences in utility sometimes result in very small differences in welfare. Since these claims are universally accepted, I don't include them as premises.

2. DOMINANCE

   If world A has higher total and average welfare and at least as much equality as world B, then A is better than B.[13]

3. LONELINESS

   If welfare is upper-bounded in lonely worlds, then it's upper-bounded in lonely individuals.

4. UNBOUNDED UTILITY

   Utility has no upper-bound for lonely individuals.

5. BOUNDED TRADEOFFS

   For any very small utility e, there's a very large utility x such that it's never worth removing x from a lonely individual in order to give e to another individual.

Prelude to the Proof

I'll briefly comment on each premise. Then I'll move to the proof.

ANTI-MONSTROSITY

Since ANTI-MONSTROSITY is simply the negation of MONSTROSITY, it's the starting point for the impossibility result. Most people find MONSTROSITY counterintuitive: it's hard to imagine that the value of a single, lonely life could outweigh the values of billions (trillions, etc…) of excellent human lives. At the end of the paper, however, I'll consider the possibility of simply accepting MONSTROSITY.

[13] DOMINANCE is slightly stronger than the principle sometimes called 'NON-ANTI-EGALITARIANISM': if world A has higher total and average welfare than world B and is perfectly equal, then A is better than B. For simplicity, I'll assume that a world with a single individual is perfectly equal, though it's worth noting that there are measures of equality that yield other verdicts.

DOMINANCE

Most population axiologies accepts DOMINANCE. This includes totalism, averagism, standard variable value views, and standard lexical threshold views. Furthermore, even population axiologies that deny the version of DOMINANCE stated above still usually accept nearby versions of the principle that suffice for nearby versions of the impossibility result.[14]

LONELINESS

Let a *lonely world* be a world that contains only a single individual, and let a *lonely individual* be an individual who is completely isolated from other individuals, meaning they lack any interactions with other individuals over the entire course of their life. Since LONELINESS is a conditional claim, it's hard to reject. If there's an upper-bound on welfare for lonely individuals who are the sole occupants of their worlds, then it's hard to see why there wouldn't also be an upper-bound on welfare for lonely individuals who are completely isolated from others within the same world.[15]

UNBOUNDED UTILITY

For any finite lifespan, it seems at least metaphysically possible for there to be an individual with a longer one. And whether the individuals under consideration are lonely or not seems irrelevant to the range of possible lifespans. If lifespans can be arbitrarily long, then UNBOUNDED UTILITY doesn't require thinking that momentary utility is unbounded. Perhaps

[14] As an example, Critical Level Totalism denies DOMINANCE: on this view, it's possible to increase average and total welfare (holding fixed inequality) while *decreasing* value by adding lives worth living that are below the critical threshold. But Critical Level Totalism still entails a nearby principle: namely, if world A has higher total and average welfare and at least as much equality as world B *and* both worlds contain only lives above the critical level, then A is better than B.

[15] In fact, LONELINESS is the only premise that strikes me as immune to attack. None of the responses I consider in §3 involve rejecting LONELINESS.

there's a limit on how much pleasure, knowledge, or desire-satisfaction one could have at any given moment. Instead, it requires only that the utility that one can accrue over the course of an entire lifetime is unbounded.

UNBOUNDED UTILITY leaves open a wide range of questions about the structure of utility. This includes whether utility is continuous or discrete, whether it's totally ordered or merely partially ordered, and whether it has a scalar or a vector structure. Furthermore, the proof makes no assumptions about the relationship between utility and welfare: for example, the proof doesn't even require assuming that increases in utility entail increases in welfare (other things being equal). For simplicity, I'll assume that utility is measurable with a ratio scale (though it's possible to develop a version of the impossibility result even if we assume that utility is measurable only with an ordinal scale). And in §3, I'll explain how to think about the impossibility result if we take utility to have a vector structure.

It's worth stating UNBOUNDED UTILITY in more formal terms. Let U be the set of utility levels accessible to lonely individuals. Then UNBOUNDED UTILITY is the claim that there's no utility level u_z such that for any other utility level $u_i \in U$, $u_z > u_i$. In other words, the set of utility values accessible to lonely individuals is unbounded from above.[16]

BOUNDED TRADEOFFS

Many think a small gain to someone with a bad life can sometimes outweigh a large loss to someone with a good life. But such intuitions are generally tempered by the background assumption that there are limits in how extreme such tradeoffs can be. To reject BOUNDED TRADEOFFS is to hold that these tradeoffs can be *arbitrarily extreme*. Imagine a loss to a utility monster and a gain to an ordinary person that you think approaches the limits of

[16] Whenever I talk about 'upper-bounds', I'll mean least upper-bounds. Note that least-upper-boundedness is distinct from having a maximal value. For example, the open interval (0, 1) is least-upper-bounded by 1 (for any $x \in (0, 1)$, $x < 1$), but it has no maximal value (there's no $z \in (0, 1)$ such that for any $x \in (0, 1)$, $x < z$). This occurs when the *ambient space* (which embeds the target set) contains an upper-bound that isn't within the target set itself.

acceptability. Now imagine that the loss is orders of magnitude worse. Perhaps, for example, the tradeoff is the loss of eons of pleasure for a gain of a mildly pleasant head-tingle for a fraction of a second. If you reject BOUNDED TRADEOFFS, it might still be better for a utility monster to receive that monstrous loss, so long as an ordinary person receives that tiny gain.

The strategy for proving the impossibility result will be to show that the first four premises contradict BOUNDED TRADEOFFS. But it's worth mentioning that the first four premises actually lead to a conclusion even worse than the negation of BOUNDED TRADEOFFS. Here's that conclusion: for any very small utility e, arbitrarily large utility x, *and* arbitrarily large finite number n, it can be worth removing x utility from n lonely individuals in order to give e utility to another individual.[17] In other words, not only can an arbitrarily small gain to an ordinary person outweigh an arbitrarily large loss to *one* utility monster, but it can outweigh arbitrarily large losses to *arbitrarily many* utility monsters. Surely that's monstrous!

The Proof

The proof is somewhat technical. Because of this, I'll start with an informal overview of the basic idea.

The basic idea is that ANTI-MONSTROSITY, alongside DOMINANCE and LONELINESS, entails an upper-bound on welfare for lonely individuals. Consequently, any lonely individual near that upper-bound can accrue at most only tiny increases in welfare. But given UNBOUNDED UTILITY, even individuals near that upper-bound will still be able to accrue massive amounts of utility. Consequently, for any arbitrarily small utility gain and arbitrarily large utility loss, we can find an ordinary person and a utility monster such that the small gain to the ordinary person can more than compensate the

[17] It would be possible to strengthen the impossibility result by replacing BOUNDED TRADEOFFS with the negation of this conclusion. This replacement premise would be even more compelling than BOUNDED TRADEOFFS. But doing this would also make the statement of and the argument for the impossibility result more complicated. So, to streamline things, I've chosen to instead formulate the impossibility result using BOUNDED TRADEOFFS.

large loss to the utility monster. But this violates BOUNDED TRADEOFFS. Hence, the five premises are mutually incompatible.

To establish that such a tradeoff can be worthwhile, the proof defines two worlds, which I'll call the 'Base World' and the 'Tradeoff World'. Both worlds contain a lonely utility monster and an ordinary human, and in both worlds the utility monster has a welfare level that's very close to the upper-bound for lonely individuals. But the utility monster in the Base World (which I'll call 'Monster+') has much more utility than the utility monster in the Tradeoff World (which I'll call 'Monster–'), while the human in the Base World has slightly less utility than the human in the Tradeoff World. The proof establishes that no matter how small the difference in utility between the humans, and no matter how large the difference in utility between the utility monsters, it can be worthwhile to make the tradeoff (violating BOUNDED TRADEOFFS).

Now let's proceed to the proof. The proof will proceed in three steps. For each step, I'll mention both what that step aims to establish, and which premises are used by that step.

Step 1: An Upper-Bound on Welfare (for Lonely Individuals)
Premises: ANTI-MONSTROSITY, DOMINANCE, LONELINESS

Suppose, for reductio, that welfare *isn't* bounded in lonely *worlds*. Then we could define a world Ω containing only a utility monster where the total welfare in Ω is greater than the total welfare in world A. Since Ω has only a single individual, it also has greater average welfare than A, and it's also perfectly equal. DOMINANCE then entails that Ω is better than A. But since A was chosen arbitrarily, it follows that for any world A containing only excellent human lives, there's a better world Ω containing only a utility monster, contradicting ANTI-MONSTROSITY.

This means that the conjunction of ANTI-MONSTROSITY and DOMINANCE entails that welfare is upper-bounded in lonely worlds. From there,

we can apply LONELINESS to yield the result that welfare is upper-bounded in lonely individuals. Let z be the least upper-bound.

Step 2: The Utility Monsters
Premise: UNBOUNDED UTILITY

Now we'll define the utility monsters. This will consist in specifying each utility monster's utility level and welfare level.

Let's start with *Monster–*, the utility monster in the Tradeoff World. To define its welfare level, we start by choosing a very small amount of utility e. Then let ε be the small amount of welfare generated by giving e to an ordinary human.[18] We'll stipulate that Monster– has welfare level w, where $w > z - \varepsilon$.[19] In other words, Monster– has a welfare level that's very close to the upper-bound for lonely individuals.

Let u be the utility level of Monster–. Note that it doesn't matter how much or how little you think u might be. And it doesn't matter if there are multiple utility levels compatible with the definition of Monster– (if that's the case, then select the lowest one). We'll simply let u denote Monster–'s utility level, whatever that might be.[20]

Now let's move on to *Monster+*, the utility monster in the Base World. Given UNBOUNDED UTILITY, for any arbitrarily large utility x, it's possible for

---

[18] If you think it's possible for an ordinary human to receive an increase in utility without an increase in welfare, then simply choose a value for e that's large enough to increase an ordinary human's welfare.

[19] This part of the proof works whether we take welfare to be discrete or continuous. Suppose first that welfare is discrete. Let y be the second-highest welfare level accessible to lonely individuals (with z being the highest). If $\varepsilon < z - y$, then it follows that the welfare of Monster– is z. On the other hand, suppose welfare is continuous. Then we can choose a welfare level w such that $z > w > z - \varepsilon$. Either way, the proof goes through.

[20] The proof makes no assumptions about the relationship between welfare and utility. More precisely, no step in the proof assumes that given some utility u, we can thereby

there to be a utility monster with utility $u+$ (= $u + x$). Define Monster+ as a lonely utility monster with utility $u+$. In other words, Monster+ has significantly more utility than Monster–.

Let w^* be the welfare level of Monster+. We'll leave open exactly how high w^* is.[21] The important point is that $w^* - w < \varepsilon$. That is, the welfare of Monster+ minus the welfare of Monster– must be less than ε. This follows from the fact that $z - w < \varepsilon$ (the welfare of Monster– is less than ε from the upper-bound for lonely individuals), alongside the fact that $w^* < z$ (the welfare of Monster+ is less than the upper-bound). Hence, $w^* - w < \varepsilon$.

To summarize, Monster– has utility u and welfare level w (which is within ε of z, the upper-bound on welfare for lonely individuals), while Monster+ has utility $u + x$ and welfare level w^* (where $w^* - w < \varepsilon$).

Step 3: The Tradeoff
Premise: DOMINANCE

Define the *Tradeoff World* as a world containing a lonely Monster– and an ordinary human with welfare level $\alpha + \varepsilon$. Define the *Base World* as a world containing a lonely Monster+ and an ordinary human with welfare level α.

Since $\alpha < (\alpha + \varepsilon)$ and $(w^* - w) < \varepsilon$, it follows that the Tradeoff World has greater average and total welfare (and greater equality) than the Base World. By DOMINANCE, the Tradeoff World is better than the Base World. Since x and e were arbitrary, this means that there's no bound to the worthwhile tradeoffs. More precisely, this means that for any very small utility e

---

determine a welfare level w. By consequence, the proof holds even for those who think that welfare is a function of more than merely utility.

[21] The proof is even compatible with the possibility that Monster– has *more* welfare than Monster+. This would be an odd view, given that Monster+ (with welfare w^*) has more utility than Monster– (with welfare w). But if $w > w^*$, then it trivially follows that $w^* - w < \varepsilon$, since $w^* - w < 0$ and $\varepsilon > 0$. Note that this also demonstrates why the impossibility result doesn't require a monotonicity principle. For example, the proof is even compatible with the view that increases in utility sometimes entail *decreases* in welfare.

(that suffices for generating a small amount of welfare in an ordinary human) and any very large utility x, it can be worth removing x from a lonely individual in order to give e to another individual. That result contradicts BOUNDED TRADEOFFS.

In fact, the result is even stronger than the negation of BOUNDED TRADEOFFS. According to BOUNDED TRADEOFFS, for *any* very small utility e, there's *some* very large utility x such that it's never worth removing x from a lonely individual in order to give e to another individual. Here's the logical form of that premise: $\forall e \exists x(\neg \mathrm{R}ex)$. The negation of BOUNDED TRADEOFFS says that for *some* very small utility e, for *every* large utility x, it can be worth removing x from a lonely individual in order to give e to another individual. Here's the logical form of the negation: $\exists e \forall x(\mathrm{R}ex)$. But the result we've established is that ANTI-MONSTROSITY, DOMINANCE, LONELINESS, and UNBOUNDED UTILITY entail that for *every* very small utility e (that results in a small change in welfare for an ordinary human) and *any* very large utility x, it can be worth removing x from a lonely individual in order to give e to another individual. Here's the logical form of that conclusion: $\forall e \forall x(\mathrm{R}ex)$.

The upshot is that the premises are mutually incompatible: it's impossible for ANTI-MONSTROSITY, DOMINANCE, LONELINESS, UNBOUNDED UTILITY, and BOUNDED TRADEOFFS to all be true.

§4 Responses to the Puzzle

I'll now turn to responses to the puzzle. For each response, I'll also note which premise of the impossibility result is being rejected.

There are some reactions to the puzzle that leave it unclear which premise, if any, is to be rejected. To indicate this unclarity, I'll mark the potentially rejected premise with '(?)'. There are also some reactions that may initially feel like a solution but that actually turn out to be compatible with all the premises—I'll discuss those as well.

Response 1: *The Values of Lives*
Rejects: DOMINANCE

Lives are valuable over and above their contribution to welfare. Therefore, even if world Ω has higher total welfare, higher average welfare, and as much equality as world B, A might still be better than B in virtue of containing a greater number of lives.[22]

To start, notice that this view must assign value to *lives*, rather than to *life*.[23] This is because the utility monster could be arbitrarily long-lived. Hence, the total quantity of life in Ω may exceed the total quantity of life in A. However, the utility monster still has only a single life. So, if each life generates some extra value, then the extra value from all the lives in A may be enough to render A better overall than Ω.

This view faces some counterintuitive consequences. First, taking lives to be valuable makes it even harder to avoid REPUGNANCE: just as A contains more lives than Ω, so too Z contains more lives than A. Second, this view entails that some worlds containing only lives not worth living can be overall good, at least if the extra value generated by the lives could outweigh the disvalue generated by the negative welfares of those lives.

Those who endorse this strategy must reject DOMINANCE: even if A has greater total and average welfare than and at least as much equality as B, B could nevertheless be better in virtue of containing more lives. But DOMINANCE was used in the proof to establish an upper-bound on welfare for lonely individuals. If there's no such upper-bound, then merely appealing to the values of lives doesn't suffice for resisting MONSTROSITY. This is because any advantage that A has over Ω with respect to number of lives could then be outweighed by giving the utility monster sufficiently high welfare.

[22] See Carter [2011] for an endorsement of this view.
[23] See Lee [2022] for argument against the idea that life itself is a welfare good.

To circumvent these worries, one could hold that the value generated by lives is *superior* to the value generated by welfare, meaning that certain increases in the number of lives outweigh *any* increases in total or average welfare.[24] A view of that kind would avoid MONSTROSITY. But such a view entails that if world A has a sufficiently large population, and world B has a sufficiently small population, then A is better than B no matter the distributions of welfare in the two worlds. In other words, even if A contains only horrible lives not worth living and B contains only excellent lives, A would still be better than B simply in virtue of the sizes of the populations. That result strikes me as more repugnant than REPUGNANCE and more monstrous than MONSTROSITY.

Response 2: *Social Goods*
Rejects: UNBOUNDED UTILITY (?)

Since the utility monster is lonely, it has no social goods. If social goods are important for welfare, then a lonely utility monster would be wholly lacking in an important component of welfare.

Many people, when first encountering MONSTROSITY, are tempted to appeal to social goods in order to resist the idea that the utility monster has an especially good life. But the impossibility result makes no assumptions about how high the upper-bound on welfare for lonely individuals might be. The proof is compatible with the view that the utility monster has at best a mediocre life (because it's lonely).

One could instead appeal to social goods to resist UNBOUNDED UTILITY. The impossibility result leaves open how to aggregate different kinds of welfare goods into overall utility levels. Perhaps the utility monster's lack of social goods places an upper-bound on the amount of *utility* that it could attain. Notice, however, that this move requires more than simply taking

[24] For discussions of superiority, see Arrhenius & Rabinowicz [2015] and Nebel [2022].

social goods to be weighted more with respect to utility than other kinds of welfare goods.[25] Instead, a rejection of UNBOUNDED UTILITY requires holding that social goods are *superior with respect to utility* to other kinds of welfare goods, meaning that there are some amounts of social goods that outweigh (with respect to utility) *any* amounts of all other goods. If that's the right way of thinking about utility, then UNBOUNDED UTILITY is false.

But even if that view were right, it would be a hollow victory. The force of the impossibility result remains: it just now needs to be restricted to the kinds of welfare goods accessible to lonely individuals. That is, it can still be worthwhile to remove arbitrarily large amounts of certain kinds of welfare goods (say, pleasure or knowledge) from a utility monster in order to give a very small amount of that same kind of welfare good to an ordinary person. Therefore, even if the utility monster's lack of social goods imposes an upper-bound on its utility simpliciter, it's still plausible that there's no upper-bound on utility for lonely individuals with respect to the kinds of welfare goods accessible to such individuals. And that restricted thesis (alongside the other premises) is enough to drive the impossibility result.

Response 3: *Multidimensionality*
Rejects: UNBOUNDED UTILITY (?)

There are multiple kinds of welfare goods, and no objective way of aggregating across different kinds of welfare goods to yield overall utility levels.

---

[25] Here's a more formal way of expressing this idea. Let g_1 and g_2 be categories of welfare goods, let $g_i(x)$ denote the amount of g_i possessed by individual x, and let $u(x)$ be the utility of x. To say that g_1 is weighted more with respect to utility than g_2 is to say that $u(x) = k \times g_1(x) + g_2(x)$, where $k > 1$. However, if lonely individuals can accrue arbitrarily large amounts of g_2, then (so long as k is finite) there will still be no upper-bound on utility for lonely individuals. Therefore, in order to reject UNBOUNDED UTILITY, one must hold that some amounts of g_1 outweigh (with respect to utility) *any* amounts of g_2.

I defined 'utility' as "overall quantity of welfare goods." But you might question whether it even makes sense to talk about such an overall quantity. Perhaps we can quantify amount of pleasure, amount of knowledge, and amount of friendship, but not the overall amount of pleasure *and* knowledge *and* friendship. Perhaps we ought to think about quantity of welfare goods multidimensionally.

However, multidimensionality isn't a reason to reject the concept of utility. Instead, it's a reason to represent utility levels using vectors (rather than scalars).[26] But taking utility to have a vector structure is compatible with UNBOUNDED UTILITY. If utility levels are vectors, then we need a principle for determining when a utility level u_1 is greater than utility level u_2. Here's a very natural criterion: $u_1 > u_2$ if (1) on some dimensions, $u_1 > u_2$, and (2) on every dimension, $u_1 \geq u_2$. In other words, we can define an ordering relation on utility levels by appealing to dominance relations with respect to values on individual dimensions of utility.

It's worth noting that the definition above is compatible with the possibility of incommensurability between utility levels.[27] Suppose, for example, that utility is three-dimensional (with the dimensions corresponding to, say, quantity of pleasure, quantity of knowledge, and quantity of friendship). Then, if x has more pleasure while y has more knowledge, there may be no fact of the matter as to whether x has more utility than y. Put another way, if utility is multidimensional, then there might be only a partial ordering (rather than a total ordering) on utility levels.

Suppose that utility is multidimensional, that some dimensions of utility (such as social goods) are inaccessible to lonely individuals, and that the partial ordering on utility levels is determined by the dominance

[26] Another possibility is to develop an account of how the dimensions aggregate to yield an overall utility value. See Hedden & Munoz [2024] on multidimensionality about value. See Hedden & Nebel [2024] on how to aggregate even when individual dimensions have different scale types.

[27] See Hájek & Rabinowicz [2022] on incommensurability with respect to value.

relation expressed above. Then the following result holds: for any utility u_i that's accessible to a lonely individual, there's another utility level u_j that's inaccessible to a lonely individual such that $u_j > u_i$.

This might initially appear in tension with UNBOUNDED UTILITY. But the result and the premise are actually compatible, since it still remains the case that there's no utility level u_z such that for any utility level u_i accessible to lonely individuals, $u_z > u_i$. The criterion for ordering utility levels defined above appealed to dominance between values along individual dimensions. But there's no utility level that dominates *every* utility level accessible to lonely individuals, since there's no upper-bound on values along the dimensions of utility that are accessible to lonely individuals. Even if lonely individuals necessarily have a zero value along some dimensions of utility, there may still be no limit to the amount of utility they could accrue.

Response 4: *Prioritarianism*

Rejects: DOMINANCE, BOUNDED TRADEOFFS

> Those who are worse-off have greater claims to receiving benefits. Therefore, it's no surprise if arbitrarily small gains to an ordinary person can outweigh arbitrarily large losses to a utility monster.

If we assume that the utility monster has a much higher welfare level than an ordinary person, then prioritarian considerations may be relevant. But the impossibility result doesn't rest on that assumption. Even if the utility monster and the ordinary person have very similar welfare levels, the argument driving the impossibility result still goes through. This is because the impossibility result leaves open the relationship between utility and welfare (including whether utility monsters have monstrous welfare levels).

Furthermore, it's possible to generate a version of the impossibility result even if we accept that utility monsters have very high welfare levels. Consider two lonely utility monsters, both of whom are very close to the

upper-bound on welfare for lonely individuals. Since both are very close to that upper-bound, they can differ only marginally in welfare. Yet, so long as we calibrate the numbers correctly, we can guarantee that for any ratio n, there's a pair of utility values u_1 and u_2 such that $u_1 = n \times u_2$, and such that it's worth conferring a loss of u_1 to one lonely utility monster in order to confer a gain of u_2 to the other lonely utility monster. In other words, the impossibility result entails that arbitrarily large differences in losses and gains can be worthwhile, even when the individuals under consideration have nearly the same welfare levels. This is because it's possible to replace the ordinary human in the impossibility result with another utility monster.

What if welfare is unbounded for lonely individuals? If we suppose this, then it's possible to appeal to prioritarian considerations to resist MONSTROSITY. This strategy is developed by Stroppa [2024], who holds that the value of a life gradually approaches an upper-bound as the life increases in welfare. More precisely, Stroppa holds that the value generated by a life is a concave, asymptotic function of the welfare of that life. On this view, utility monsters indeed have monstrous welfare levels, but the value of a utility monster's life is only modestly higher than the value of an excellent human life. This view resists MONSTROSITY: world Ω (containing a lonely utility monster) has at most the value equal to the upper-bound on value for a single life, whereas world A (containing many excellent human lives) has much greater value. But this view faces some consequences that many will find even costlier.

First, the view denies DOMINANCE: even if Ω has higher total and average welfare than (and, trivially, at least as much equality as) world A, Ω might still be worse than A (since there's an upper-bound on the amount of value that may be contributed by any single life). Moreover, its denial of DOMINANCE isn't motivated by appeal to another contributor to value besides welfare and equality (for example, the values of lives). Instead, the view denies DOMINANCE despite taking welfare to be the only factor that determines value. Second, the view has to not only deny BOUNDED TRADEOFFS, but also the corresponding principle that's framed in terms of

welfare (instead of utility) and that involves harms to arbitrarily many utility monsters (instead of just one). In other words, the view entails that an arbitrarily small benefit to a human can outweigh arbitrarily large harms to arbitrarily many utility monsters. That's a hard bullet to bite: I suspect even most with prioritarian sympathies would be reluctant to accept that such tradeoffs can be arbitrarily extreme at arbitrarily large scales.

Response 5: *Nomic Limits*
Rejects: UNBOUNDED UTILITY

The laws of nature limit how long lifespans could be. And if there are limits on lifespans, then there are limits on the amount of utility that could be accrued within a single life.

When dealing with axiological puzzles, it's reasonable to abstract away from nomic limits. To illustrate, let n be the *carrying capacity* of the world, meaning that it's nomically impossible for the world to contain more than n individuals. Now consider REPUGNANCE: for any world A containing only excellent lives, there's a better world Z containing a much larger population consisting only of lives barely worth living. Suppose that at each step in the spectrum argument for REPUGNANCE, the number of lives in the population is multiplied by a factor of 2 (so, for example, world B contains twice as many lives as world A). And imagine that there's a world A containing only excellent lives that's near enough to the carrying capacity (say, at 0.1% of the carrying capacity) such that there's no nomically possible world Z that contains a much larger population consisting only of lives barely worth living that has greater total welfare.[28] Have we now discovered a new solution to the most prominent problem in population ethics?

[28] At these numbers, it takes only 10 steps to move from A to the carrying capacity, since $0.1 \times 2^{10} > 100$. So, if it takes more than 10 steps to move from world A to world Z, it follows that for any A-world at 0.1% (or more) of the carrying capacity, there's no better Z-world.

Few would find this to be a satisfactory response to REPUGNANCE. Even if the claims about nomic limits are correct, the response tells us little about which population axiology is correct. To deal with REPUGNANCE, one must develop and defend a population axiology. Similarly, to deal with MONSTROSITY, one must combine a plausible population axiology with a plausible welfare axiology. Both MONSTROSITY and REPUGNANCE are axiological puzzles. Therefore, both warrant axiological solutions.

Response 6: *Personal Identity*
Rejects: UNBOUNDED UTILITY

Any sufficiently long life would no longer be the life of a single person; early person-stages of such a life wouldn't bear the relevant psychological or physiological relations to later person-stages of that life.

It's not clear that it's *impossible* for individuals to live arbitrarily long lives. On the most prominent theories of personal identity, numerical identity between persons is a matter of continuity and/or connectedness along psychological and/or physiological dimensions. There's no obvious reason for thinking that the psychology of the utility monster must radically change over time, and we can stipulate that the utility monster's mind and body vary only gradually over any given interval of time.

Furthermore, a number of philosophical discussions take for granted the possibility of extremely long lives. This is most obvious in the literature on immortality, but also in some discussions of the values of very long lives.[29] Since the possibility of very long lives is compatible with standard theories of personal identity and commonly assumed in some philosophical discussions, it's not enough to simply assert that the utility monster

[29] See Williams [1973] for a classic discussion of immortality, and Greene [2017] for a recent discussion arguing that very long (possibly infinitely long) lives could be very good lives. See Olson [2023] for a recent overview of theories of personal identity.

couldn't be a single person. Instead, developing this objection would require one to motivate a theory of personal identity that precludes the possibility of extremely long-lived individuals.[30]

Response 7: *Monstrosity*
Rejects: ANTI-MONSTROSITY

Accept MONSTROSITY: for any world A containing only excellent human lives, there's a better world Ω containing only a lone utility monster.

I favor this response; I think the best way to develop a satisfactory axiology is to simply accept MONSTROSITY. Since many will find this counterintuitive, I'll make a few remarks about what sways me. The following lines of argument aren't original, and I don't expect them to convince the skeptic.

First, initial intuitions may sometimes be unreliable. Perhaps we're psychologically incapable of grasping the overall goodness of the utility monster's life. And perhaps we tacitly assume that the metaphysical bounds of welfare only marginally exceed the bounds accessible to humans.[31] But imagine that, say, snails are welfare subjects (yet capable of only slightly positive welfare levels), and compare a human life to a large population of snail lives. It strikes me as reasonable to think that one excellent human life could outweigh many—even millions—of snail lives. Perhaps there could be creatures that stand to us as we stand to snails.

Second, many of us have become persuaded that the boundaries between persons are less metaphysically significant than we might have

[30] Some, such as Lewis [1976], maintain that even a life that only gradually changes (and is very long) cannot be the life of a single person. I think there are ways of reshaping the puzzle of this paper to accommodate even these views. But since these reshapings would take us astray from the main dialectic, I won't further pursue this line of inquiry.

[31] See Tännsjö [2002], Huemer [2008], and Gustaffson [2022] on some similar points regarding REPUGNANCE.

initially thought.[32] If that's correct, then there may be no deep metaphysical fact of the matter about whether the utility monster is a single person living a very long life or many individuals whose lives are seamlessly connected. If that difference is metaphysically unimportant, then it may also be ethically unimportant. And if we were to think of the utility monster as a population of individuals (rather than a single individual), many would find the initial intuitions against MONSTROSITY less compelling.

Third, a number of authors have shown that developing an axiology that satisfies all commonly held intuitions is not merely difficult, but impossible.[33] This means that any axiology must relinquish some claims that strike many as intuitive. If we accept MONSTROSITY, then we free ourselves to develop an axiology around other principles that we might regard as more axiomatic. If we reject MONSTROSITY, then we'll have to pay the price through counterintuitive consequences that arise elsewhere in the theory.

Conclusion

Although the utility monster is a famous character in philosophical lore, there's surprisingly little discussion of it in the contemporary literature. This may be because it's often dismissed as an idle thought-experiment or a metaphysical impossibility. But one of the main innovations of this paper has been to distinguish utility—defined as quantity of welfare goods—from welfare—the benefit that an individual receives from those welfare goods. It's plausible that there could be an individual who accrues monstrous quantities of welfare goods over the course of a very long lifetime. And if we grant that possibility, then a new philosophical puzzle arises.

[32] See Parfit [1984] for many of the classic arguments for this sort of view. See Hedden [2015] for arguments in favor of treating time-slices of people as the locus of normativity.

[33] See Ng [1989], Carlson [1998], Arrhenius [2000], Kitcher [2000] for some well-known impossibility results concerning REPUGNANCE. In light of this, many ethicists (see Zuber *et al* [2021]) now think that avoidance of REPUGNANCE ought not be taken as mandatory for population axiologies.

To solve that puzzle, we need to combine a population axiology with a welfare axiology. Just as we can ask how distributions of welfare relate to the overall values of worlds, so too we can ask how distributions of utility relate to the overall welfares of lives. Although there are obvious connections between population ethics and theories of welfare, there has been relatively little systematic exploration of the intersection of these subject-matters. The puzzle of MONSTROSITY forces us to think carefully about how these areas interact.

The impossibility result generates a set of choice points for future axiologies. Which premise should we reject? Some may question UNBOUNDED UTILITY, perhaps by appealing to multidimensionality or to principled limits on lengths of lives. Some may think there are independent reasons for resisting DOMINANCE, perhaps due to the values of lives. Some may think we can live with rejecting BOUNDED TRADEOFFS, perhaps as an extension of prioritarian motivations. And some, like me, may wish to embrace the utility monster by rejecting ANTI-MONSTROSITY.

But whichever premise we reject, we'll have to give up some principles that strike many as intuitive. And as we've learned with REPUGNANCE, it may be impossible to satisfy all the intuitions that we started with. Therefore, I suspect that any future axiology must—in one way or another—be monstrous.

References

Arrhenius, G. (2000). An impossibility theorem for welfarist axiologies. *Economics & Philosophy*, 16(2), 247-266

Arrhenius, G., & Rabinowicz, W. (2015). Value superiority. The Oxford handbook of value theory, 225-248.

Arrhenius, Gustaf & Andersson, Emil (2021). The Repugnant Conclusion: An Overview. In Stephen M. Gardiner (ed.), The Oxford Handbook of Intergenerational Ethics. Oxford University Press.

Blackorby, C., Bossert, W., & Donaldson, D. (1997). Critical-level utilitarianism and the population-ethics dilemma. Economics & Philosophy, 13(2), 197-230.

Broome, J. (2004). Weighing lives. OUP Oxford University Press.

Carlson, E. (1998). Mere addition and two trilemmas of population ethics. Economics and Philosophy, 14 , 283–306.

Carter, A. (2011). Some groundwork for a multidimensional axiology. *Philosophical studies*, *154*, 389-408

Greaves, H. (2017). Population axiology. *Philosophy Compass*, *12*(11), e12442. https://doi.org/10.1111/phc3.12442

Greene, Preston (2017). Value in Very Long Lives. *Journal of Moral Philosophy* 14 (4):416-434.

Gustafsson, Johan E. 2022. Our Intuitive Grasp of the Repugnant Conclusion, in The Oxford Handbook of Population Ethics (New York: Oxford University Press).

Hájek, A., & Rabinowicz, W. (2022). Degrees of commensurability and the repugnant conclusion. Noûs, 56(4), 897-919.

Hedden, Brian (2015). Reasons Without Persons: Rationality, Identity, and Time. Oxford, United Kingdom: Oxford University Press UK.

Hedden, Brian & Muñoz, Daniel (2024). Dimensions of Value. Noûs 58 (2):291-305.

Hedden, Brian & Nebel, Jacob M. (2024). Multidimensional Concepts and Disparate Scale Types. Philosophical Review.

Huemer, Michael. 2008. In Defence of Repugnance, Mind 117.468: 899–933.

Hurka, T. (1983). Value and population size. Ethics, 93(3), 496-507.

Kitcher, P. (2000). Parfit's Puzzle. Nous, 34 (4), 550–577. doi: 10.1111/0029- 4624.00278

Lee, Andrew Y. (2023). The Neutrality of Life. Australasian Journal of Philosophy:1-19.

Lewis, David (1976). I Survival and Identity. In Amélie Oksenberg Rorty (ed.), Identities of Persons. University of California Press. pp. 17-40.

Nebel, J. M.. (2022) Totalism without Repugnance. In McMahan, J., Campbell, T., Goodrich, J., & Ramakrishnan, K. (eds), Ethics and Existence: The Legacy of Derek Parfit, 200-31.

Ng, Y. K. (1989). What Should We Do About Future Generations?: Impossibility of Parfit's Theory X. Economics & Philosophy, 5(2), 235-253.

Nozick, R. (1974). Anarchy, state, and utopia. New York: Basic Books.

Olson, Eric T., "Personal Identity", *The Stanford Encyclopedia of Philosophy* (Fall 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/identity-personal/>.

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

Pivato, M. (2020). Rank-additive population ethics. *Economic Theory*, *69*(4), 861-918.

Portmore, D. W., 1999, "Does the Total Principle Have Any Repugnant Implications?" *Ratio*, 12 (1): 80–98.

Sen, A. (1977). On weights and measures: informational constraints in social welfare analysis. *Econometrica: Journal of the Econometric Society*, 1539-1572.

Sider, T. R. (1991). Might theory X be a theory of diminishing marginal value?. *Analysis*, *51*(4), 265-271.

Stroppa, Luca (2024). The Monstrous Conclusion. Synthese 203 (6):1-24.

Tännsjö, Torbjörn. 2002. Why We Ought to Accept the Repugnant Conclusion, Utilitas, 14.3: 339–59.

Thomas, Teruji (2018). Some Possibilities in Population Axiology. Mind 127 (507):807-832.

Williams Bernard (1973), 'The Makropolous Case: Reflections on the Tedium of Immortality', in B. Williams, *Problems of the Self: Philosophical Papers 1956–1972* (Cambridge University Press), pp. 82–100.

Zuber, Stephane ; Venkatesh, Nikhil ; Tännsjö, Torbjörn ; Tarsney, Christian ; Stefánsson, H. Orri ; Steele, Katie ; Spears, Dean ; Sebo, Jeff ; Pivato, Marcus ; Ord, Toby ; Ng, Yew-Kwang ; Masny, Michal ; MacAskill, William ; Lawson, Nicholas ; Kuruc, Kevin ; Hutchinson, Michelle ; Gustafsson, Johan E. ; Greaves, Hilary ; Forsberg, Lisa ; Fleurbaey, Marc ; Coffey, Diane ; Cato, Susumu ; Castro, Clinton ; Campbell, Tim ; Budolfson, Mark ; Broome, John ; Berger, Alexander ; Beckstead, Nick & Asheim, Geir B. (2021). What Should We Agree on about the Repugnant Conclusion? Utilitas 33 (4):379-383.