Negative Utility Monsters*

Richard Yetter Chappell

(Forthcoming in *Utilitas*)

Abstract

Many consider Nozick's "utility monster"—a being more efficient than ordinary people at converting resources into well-being, with no upper limit—to be a damning counterexample to utilitarianism. But our intuitions may be reversed by considering a variation in which the utility monster starts from a baseline status of massive suffering. This suggests a rethinking of the force of the original objection.

Introduction

Nozick (1974, 41) famously objected that "Utilitarian theory is embarrassed by the possibility of utility monsters who get enormously greater gains of utility from any sacrifice of others than these others lose. For, unacceptably, the theory seems to require that we all be sacrificed in the monster's maw, in order to increase total utility."

*Thanks to Theron Pummer, Helen Yetter-Chappell, and anonymous referees, for helpful comments.

After isolating the distinctive feature of this objection (in contrast to standard demandingness and rights-sacrifice objections), I show how it can be undermined by considering a variation in which the utility monster starts from a position of massive suffering. I close by considering the implications for the original objection.

1 Isolating the Objection

The utility monster scenario may seem intuitively objectionable for many reasons, but only one is the intended target of this paper. Some may object to any general requirement of sacrifice to help others, however deserving those others may be. But this general *demandingness objection* to utilitarianism is not my target here. Others may object to the putative rights violations involved in harming some to benefit others (again, no matter how deserving those others might be). Such general concerns about utilitarian sacrifice are not my target here. Beyond these familiar objections (which might as well be illustrated by any number of other possible examples), I take there to be a further objection in the running which is more distinctively supported by the utility monster sce-

nario in particular. This objection draws on the nature of the utilitarian beneficiary: that in this case, it is just one individual, set against everyone else in the world. The *utility monster objection*, as I understand it, thus rests upon the apparent absurdity of allowing a single individual's interests to trump everyone else's.

A standard response to this objection is to question whether the utility monster scenario is really coherent (Parfit 1984, 389). How well-off (in terms of wellbeing, not resources) can a single individual be? We may plausibly hold there to be a cap or upper bound on how high one's wellbeing can go, such that benefits to one (starting from a neutral baseline) simply cannot be sufficiently large to outweigh great harms to a great many.

I, for one, cannot positively conceive of such a high level of wellbeing as to render sacrificing all to Nozick's monster a genuinely utility-maximizing act. And I doubt that I am unusual in this; I expect such imaginative resistance to the scenario to be commonplace. If so, that would seem to suffice to explain our intuitive aversion to sacrificing all to the monster, without necessarily undermining utiltiarian theory at all.

2 The Negative Utility Monster

We may attempt to restore coherence to the utility monster scenario in two steps. First, we allow the benefits to the monster to be spread out over time. Perhaps each sacrifice we make to the monster gives it another century of maximally good life, for example. This might already be enough for some to think that benefiting the monster isn't such an obviously wrong option. But I don't think it suffices, as there are strong intuitive grounds for denying a simple additive view of how additional good life contributes to one's lifetime wellbeing.

Consider: it would seem prudentially irrational to give up a guaranteed fifty additional years of good life for a 50% chance of one hundred additional good years (and 50% chance of instant death), even assuming no debilitation from aging. One hundred good years for an individual intuitively isn't twice as valuable as fifty. Why not? One possible explanation is that a large component of our lifetime wellbeing is determined by certain core projects that can be fully achieved within a normal lifespan. Ensuring that one's life is not too short to achieve such core projects can thus make a

huge difference to one's lifetime wellbeing, whereas any period of additional years beyond what's needed is relegated to the status of a minor bonus.

If that's right, then we cannot secure massive welfare gains for the utility monster just by massively increasing the quantity of good life that they experience. To fully fix the thought experiment, I propose that, besides spreading out benefits over time, we additionally shift the monster's *baseline* welfare level. For, as Parfit (1984 chp. 18) noted, the badness of aggregate suffering cannot plausibly be capped. As a result, we may make the monster's baseline wellbeing level as deeply negative as you care to imagine, just by imagining him to be arbitrarily long-lived, unkillable, and suffering immensely at every moment that he lives. There is now the potential for the interests of this one "Negative Utility Monster"—call him "NUM"—to really be sufficiently great in magnitude as to outweigh (in utilitarian terms) the interests of all us ordinary mortals. It becomes less clear that utilitarians need feel "embarrassed" about their verdict in this case, however.

To set up the case most neatly (avoiding confounding intuitions about demandingness, rights, and so forth), let us restrict our

focus to the question of how we ought to allocate some great pile of antecedently *unallocated* resources. There are, of course, a great many people who could benefit from having more resources. But suppose that, in each instance, the marginal benefit to NUM of granting him the additional resource (in terms of reducing his suffering and even allowing his life to contain some positive moments in their place) would far outweigh the gains anyone else could get from the resource in question. Any resource that might provide a week of relief from mild suffering for a human could instead provide a year of relief from torturous agony for NUM, let's say. The utilitarian verdict is, then, that we should give all the resources to NUM. But this does not strike me as an embarrassing verdict at all. Indeed, it strikes me as very plausibly *correct*.

This doesn't suffice to defend utilitarianism against all possible objections, of course. The restriction to unallocated resources was precisely designed to sidestep some of the most pressing intuitive objections to utilitarian sacrifice. But whatever other objections one might have, the present discussion should at least serve to undermine the distinctive force of the utility monster objection, understood (as above) as suggesting that a single individ-

ual's interests—however great—should not be allowed to outweigh everyone else's combined. For it seems, in the above case, that NUM's interests do and should outweigh all others.

To make the case more awkward for utilitarians, suppose we remove our restriction to unallocated resources, and raise the further question whether resources previously held by others should be *redistributed* to NUM to give him further relief. Again, we are to suppose that the relief he gains far outweighs the harm done to those who are newly deprived of their resources (even if they die as a result). To cancel any complications stemming from our instrumental value to future generations, suppose it is guaranteed that there will be no future generations in any case. Humanity can either enjoy itself for a last few years before collapsing, or we may end ourselves prematurely in service of relieving NUM's remaining suffering. In this case, I grant that the utilitarian verdict—that we must all sacrifice ourselves to NUM—is much less obviously correct (there are, by design, reasonable grounds for objection that were missing from the previous case). But it remains, I believe, a perfectly defensible—and entirely unembarrassing—verdict.

Utilitarians may, for example, reasonably judge their critics here

to be influenced by an unjust status-quo bias: unjustifiably favouring those of us in a privileged starting position relative to poor suffering NUM. Why should NUM have to suffer so, just because the resources he so needs are initially to be found in our possession? To lose my life would of course be a great cost to me, but not nearly so great as the centuries of torturous agony that would otherwise be suffered by NUM. So it would seem unsurprising (and certainly no cause for embarrassment) for an impartial moral view to judge NUM's interests here to be of greater moral weight than my own.

3 Implications

We've found that the Negative Utility Monster seems less intuitively threatening to utilitarianism than Nozick's original monster did. What can we learn from this? One immediate upshot, I've suggested, is to undermine the original objection, for NUM shows us that there's nothing *necessarily* objectionable about having the interests of one individual outweigh all others. But why, then, did Nozick's case seem so damning? I see two possible explanations.

The first explanation is that we are simply misled by an unwitting divergence between the arguably incoherent theoretical stipulations of the original utility monster scenario (as involving an unbounded capacity for positive utility) and the (capped-utility) scenario that we actually end up imagining. If any creature that you imagine necessarily has a modest upper bound on how well-off it can possibly get, then *of course* it would be terribly wrong to sacrifice all others merely to make this one individual a bit more happy than he already was. It would also be terribly *bad*, in terms of utility or net welfare. Our intuitive judgment about the capped-utility scenario is thus not in conflict with utilitarianism. And if the alternative, of unbounded positive utility, is indeed incoherent or otherwise unimaginable for us, then this capped-utility scenario is the only one we can bring to mind when prompted to make an intuitive judgment about Nozick's utility monster. That is, we correctly judge that it'd be a terrible mistake to feed everyone to the monster we imagine upon reading Nozick's thought experiment. Our error is to assume that the monster we have imagined is one that matches Nozick's stipulations, such that utility would really be maximized by sacrificing everyone else.

Some readers may find the above explanation tendentious, however, as it crucially relies upon the assumption that we cannot really imagine a positive utility monster at all. Some readers may be inclined to insist that, whatever imaginative blocks others of us might face, *they* can imagine it perfectly well. That is, they can imagine a creature such that it would be transparently good (in terms of utility) to sacrifice all others to it. And when they imagine this, it nonetheless strikes them as a morally bad outcome. What can be said to one who takes this view?

If they share my sense that utilitarianism yields plausible verdicts regarding the *negative* utility monster, such a defender of the Nozickian monster may naturally wonder what the relevant difference between the two cases is. They reject my first explanation—that the Nozickian monster is incoherent or otherwise unimaginable. So let me offer an alternative suggestion.

The central difference between the two cases is the monster's baseline level of wellbeing. A presumed neutral (or positive) baseline affords opportunities to boost the monster's happiness (as in Nozick's case), whereas the negative starting condition of NUM means that increments to his welfare instead take the form of relieving

or offsetting suffering. This is all very suggestive of the standard prioritarian intuition that benefits to an individual matter more the worse-off that individual is (Parfit 1997). On a prioritarian account, it will be very difficult to justify greatly harming or sacrificing people merely to provide benefits to others who are already reasonably well-off (let alone to a single such individual). NUM, by contrast, is the worst-off individual in existence, and so has high priority given to his interests when we are in a position to aid him.[1]

It thus seems that the utility monster scenario is really just pumping standard prioritarian intuitions, rather than providing the basis for a distinctively compelling objection to utilitarianism in its own right. This result, too, is arguably less troubling for utilitarians than the supposedly damning objection that we began with. Utilitarians will already have something to say about prioritarianism. Some may consider it a sufficiently minor variant on their own view that they are not concerned to dispute it. Others may be happy to demote our prioritarian judgments to the status

[1]Though, as Pummer (n.d.) notes, prioritarians may face distinctive "Priority Monster" problems if an individual like NUM is *so* much worse-off than the rest of us that the slightest relief to them is allowed to justify imposing great harms upon everyone else.

of useful heuristics for promoting utility in the face of (e.g.) the diminishing marginal utility of resources and the greater scope for improvement when we focus on the worse-off (cf. Greene and Baron 2001). Either way, assimilating the utility monster to the more familiar challenge of prioritarian intuitions should prove a comforting result for utilitarians.

One important proviso: there remains room for one to hold that it's better to distribute benefits broadly, even when this goes against prioritarianism. Suppose that NUM is at -220 wellbeing, and nine other individuals are just slightly better-off, at -200. Further suppose that we have ten resources to distribute, each of which could either relieve two points of suffering for NUM, or one for any other individual. Some people may prefer to distribute the resources equally (yielding -218 wellbeing for NUM and -199 for the other nine) rather than giving all the resources to NUM to equalize wellbeing at -200.[2]

In response to such intuitions, I would want to hear more about how such a verdict could be justified in principle—why regard a broader distribution of benefits to be fairer or better when an-

[2]Thanks to an anonymous referee for suggesting this case.

tecedent inequalities meant that some had greater need? Identical treatment may be what's fair when all involved have identical interests, but when this background condition isn't met it would instead seem fundamentally unfair to treat those with greater needs (and greater capacity to benefit from intervention) no differently than those who are already better-off.

So, I would dispute the critic's proposed verdict in this case. But perhaps I'm wrong about that. Even so, the critic's intuitive judgment here seems likely to be, at best, highly tentative. Utilitarians may judge the case differently, without embarrassment. So even if people can reasonably disagree with utilitarian verdicts in cases that set the interests of one against many (the distinctive feature of the utility monster case), this does not seem to provide the basis for a decisive or even particularly forceful objection. The intuitive force of Nozick's original case is better accounted for via my two earlier explanations.

4 Conclusion

Nozick's utility monster should no longer be seen as a damning objection to utilitarianism. The intuitive force of the case is undermined by considering a variant with immensely negative wellbeing. Offering significant relief to such a "Negative Utility Monster" plausibly *should* outweigh smaller harms or benefits to others. Our diverging intuitions about the two kinds of utility monsters may be explained conservatively as involving standard prioritarian intuitions: holding that benefits matter more the worse-off their recipient is (and matter less, the better-off their recipient is). This verdict undermines the distinctiveness of the utility monster objection, and reduces its force to whatever level one attributes to prioritarian intuitions in general. More ambitiously, the divergence between the two cases may be taken to support attempts to entirely explain away the original utility-monster intuition, e.g. as illicitly neglecting the existence of an upper bound on the monster's wellbeing. Such an explanation, if successful, suggests that our intuition about the original utility monster scenario was based on a mistake. Either way, the force of Nozick's objection is

significantly undermined by the Negative Utility Monster.

References

Greene, Joshua, and Jonathan Baron. 2001. "Intuitions About Declining Marginal Utility." *Journal of Behavioral Decision Making* 14: 243–55.

Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books.

Parfit, Derek. 1984. *Reasons and Persons*. 1987 reprint. New York: Oxford University Press.

———. 1997. "Equality and Priority." *Ratio* 10 (3): 202–21.

Pummer, Theron. n.d. "The Priority Monster."