MERE ADDITION AND THE SEPARATENESS OF PERSONS: A CANDIDATE FOR THEORY X[*]

Paper prepared for the 12th meeting of the International Society for Utilitarian Studies, New York, 8-11 August 2012

Matthew Rendall
Lecturer
School of Politics and International Relations
University of Nottingham
School of Politics and International Relations
University Park NG7 2RD
United Kingdom
44-(0)115-846-6231 (tel.)
44-(0)115-951-4859 (fax)
Matthew.Rendall@nottingham.ac.uk

Abstract: The best reason to preserve life on earth is that more lives worth living are better than fewer. But how shall we avoid the repugnant conclusion? Some suggest that we can block it by claiming that good lives are lexically better than mediocre ones. Yet this does not solve the mere addition paradox. I argue that the paradox arises because it entails two types of aggregation.

Suppose we are comparing a single good life in world J and a million sweatshop lives in world K. Only J's welfare falls above the lexical level. Since everyone in either world will be equally well-off, we adopt, in Rawls's phrase, 'the principle of rational choice for one man'. All things considered, we would prefer the single good life in J to a million sweatshop lives, and so would any other possible people. But after possible lives have become actual——when comparing J+ and K—we must weigh one person's preference for his good life against a *million other people's* preferences for moving on to K. These inevitably outweigh the preferences of the original person, making K better than J+.

If J is better than K, and K is better than J+, then either J+ must be worse than J, or we must accept intransitivity of value. I argue that J+ is worse, because of its inequality. But this is not because inequality is bad in itself. Rather, it is because J+'s inequality makes it better to move on to K. Since K is worse than J, and J+ has a property that makes it better to choose this inferior outcome, that makes J+ worse than J. This solves the mere addition paradox and lays the foundation for an impersonal theory of population ethics.

[*] The author thanks audiences at the universities of Uppsala and Zurich for comments on earlier drafts—particularly Dominic Roser and Anton Leist—and the Centre for Ethics at the University of Zurich for funding my attendance at its 2011 workshop on the non-identity problem.

I. THE BADNESS OF EXTINCTION

In 2004 Richard Posner estimated the cost of human extinction—'*very* conservatively'—at $600 trillion. Many would say it was higher than that. A nuclear war that killed 100 percent of the world's population, Derek Parfit maintains, would be far worse than a nuclear war that killed 99 percent. On one estimate it would be 'something like a million times worse…in terms of the number of people who would thereby never live'.[1] Why should anyone think these things? The most obvious reason is all the meaningful, enjoyable lives extinction would foreclose. But this appears to imply the repugnant conclusion. If we should preserve life because it maximizes total utility, then the best population could be a vast multitude, all of whose lives are barely worth living, if only there are enough of them.

Imagine a continent inhabited by ten billion people, all equally very well off, and call this state of affairs A. Unknown to the people of the first continent, you may choose to populate a second continent with a second group of the same size, all of whom will be worse off than the first group, but whose lives will still be worth living. This combined population, A+, seems as good as A, if not better. Suppose that then that you can redistribute between the continents—still without their inhabitants learning of each other's existence—so that you reduce each person's utility on the first continent, but raise the utility of each on the second continent by a greater amount. The resulting population, B, contains 20 billion people, all of whose lives are of equally high quality, but not quite as good as those of the people in A or the better-off group in A+. Again, moving from A+ to B seems an improvement, on grounds of both increased total utility and equality. Finally, imagine that you repeat the process twenty-four times. Each step seems to lead to an outcome that is better than the preceding

[1] R. A. Posner, *Catastrophe: Risk and Response* (Oxford UP, 2004), p. 141, emphasis in original; D. Parfit, *Reasons and Persons* (Oxford: Clarendon Press, 1987), p. 453; C. Sagan and R. Turco, *A Path Where No Man Thought: Nuclear Winter and the End of the Arms Race* (London: Century, 1990), p. 72.

one. The result is a huge population, Z, whose lives are just barely worth living. Could this be the best of possible worlds?[2]

Most of us find it easy to believe that Z is worse. Some believe that while adding any life worth living is good, adding lives *well* worth living is better—so much better that no number of mediocre lives outweigh a single good one. What is harder to see is how it could make the world worse to move from A to A+. The mere addition of the new people harms no one, and adds worthwhile lives. Indeed, it is hard to see why creating lives worth living should not be *good,* if it has no effect on others. As Michael Huemer says, 'Worthwhile lives are good. More of a good thing is better'.[3] But once we have added the new people, it seems equally hard to say that we should not redistribute—which both increases utility and benefits the worse-off—and move to B. Having reached B, there seems no reason not to move to B+…and on down the alphabet. Showing why A could be better than Z is not enough to block the repugnant conclusion—we need to solve the mere addition paradox.

This paper maintains that A+ is better than A, and B is better than A+. Nevertheless, at some point—perhaps at letter J—moving further down the sequence brings the population below a lexical threshold. At that point, we should stop: J+ is *not* better than J. The reason is J+'s inequality. Many authors have suggested that the inequality after mere addition is somehow bad. The challenge is to explain how adding worthwhile lives could be bad all things considered—or, for that matter, how inequality that harms no one could be bad at all.[4] Here I argue that the inequality in J+ is bad for a simple reason: it makes it better to move on to K.

For four decades 'the separateness of persons' has been a central theme of moral philosophy. The tendency of utilitarianism, Rawls famously argued, is 'to adopt for society as

[2] Parfit, *Reasons and Persons*, part IV.
[3] M. Huemer, 'In Defence of Repugnance', *Mind*, 117 (2008), pp. 899-933, at p. 923.
[4] Parfit, *Reasons and Persons*, 425; D. Parfit, 'Equality or Priority?' in M. Clayton and A. Williams (eds) *The Ideal of Equality* (Basingstoke: Palgrave Macmillan, 2002), pp. 81-125.

a whole the principle of rational choice for one man'.[5] A popular approach is to ask oneself which set of lives one would prefer if one could live them all end to end. We could plausibly prefer a single good life to any number of poor ones. This method works well so long as everyone in both outcomes we are comparing will be equally well-off. But even on utilitarian grounds some trade-offs that maximize value within a single life do not do so when some people enjoy the benefits and others bear the burdens. Notably, one can give *lexical* priority to goods within one's own life when one should not do so in interpersonal cases.

In the choice between worlds J and K, whichever world we choose, all the parties who ever exist will be equally well-off. Given a choice between a single good life in world J, and all the lives in world K lived end to end, we would choose the former. Possible people, choosing for themselves, would reason the same way. We thus consider J lexically superior. But this is not the choice we face after mere addition. Now, if we refuse to redistribute, we will *leave some better off at the expense of others.* To say 'I would rather have a J life than any number of K lives' is one thing; to say 'I would rather have a superior life in J+ than for any number of *you* to have improved lives in K' quite another. Once possible lives become actual, their aggregated preferences for redistribution will outweigh the preference of the original group for remaining in J+, making it best to move to K. But if J+ has a property—inequality—that makes it better to move to an outcome that is worse than J, then J+ itself is worse. By avoiding the repugnant conclusion and solving the mere addition paradox, this clears the way for an impersonal population theory.

II. THE REPUGNANT CONCLUSION AND INTRAPERSONAL AGGREGATION

Could we swallow the repugnant conclusion? C. L. Lewis proposed comparing the value of two worlds by asking in which we would rather live through the lives from

[5] J. Rawls, *A Theory of Justice: Revised Edition* (Cambridge, MA: Belknap Press, 1999), p. 24.

beginning to end.[6] Suppose we are offered either a hundred wonderful years, or an eternity that is barely worth living—spent, say, on a life-support machine watching TV, *bad* TV. Most of us, as Parfit argues, would pick the Century of Ecstasy.[7] Why? Some might think that the pleasures of Z are lexically inferior. Great sex and sitcom re-runs are both better than nothing, but no amount of the second could be as good as the first. This view is implausible. Great sex may be better than great movies, but is a *minute* of great sex better than *ten years* of great movies? Great movies are better than great TV, but is one minute of a great movie better than ten years of great TV? But if enough of an inferior good can outweigh a superior good, we will proceed down the alphabet to eternal shopping channel. Moreover, we can change the example so that you can have your cake and eat it too. Suppose once every century, someone will come into your room, switch off the TV, and have a minute of great sex with you. In a few eons, you will enjoy more sex *and* television than anyone who has ever lived. Yet this life seems nearly as drab.[8]

The reason to choose the century of ecstasy cannot be that it contains more total pleasure. Rather, it seems that we judge it better *overall*. On a hedonic account of well-being this makes little sense, but on a desire-utilitarian account it is plausible. '[T]he relevant notion of aggregation', James Griffin argues, 'cannot be simply that of summing up small utilities from local satisfactions; the structure of desires already incorporates, constitutes, aggregation'. Our judgments of the value of lives 'take a global form: this way of living, all in all, is better than that'.[9] To claim that *some* of one good can be better than *any* amount of

[6] C. L. Lewis, *An Analysis of Knowledge and Valuation* (La Salle: Open Court, 1946), pp. 546-47, 550-51.
[7] D. Parfit, 'Overpopulation and the Quality of Life', in P. Singer (ed.) *Applied Ethics* (Oxford UP, 1986), pp. 145-64, at p. 160; for the TV example, J. Skorupski, 'Value and Distribution', in Skorupski, *Ethical Explorations,* (Oxford UP, 1999), pp. 85-106, at p. 94.
[8] I. Persson, 'The Root of the Repugnant Conclusion and Its Rebuttal', in J. Ryberg and T. Tännsjö (eds.), *The Repugnant Conclusion: Essays on Population Ethics* (Dordrecht: Kluwer, 2004), pp. 187-99, at p. 190; J. Ryberg, 'Parfit's Repugnant Conclusion', *Philosophical Quarterly*, 46 (1996), pp. 202-13.
[9] J. Griffin, *Well-Being: Its Meaning, Measurement, and Moral Importance* (Oxford UP, 1986), pp. 15, 34-35.

another is not believable if one is comparing pleasures, but is plausible when the first of the goods is a successful life. One successful life beats any number of failures.[10]

Lexical views make some writers see red. They imply, says Michael Huemer, that 'enabling a few people to hear Mozart's music might be more important than providing food, shelter, and medical care to millions.[11] But in A, by stipulation, everyone is very well off; there *are* no hungry or homeless people. If we call A better than Z, we are not saying that it is better to neglect the badly-off. We are saying that it is better to have a smaller population, *all* of whose members are very well off. Huemer's real worry seems to be that we could add needy people to this population and then claim that redistribution would make the world worse. I am going to argue that to do *that* would be very wrong indeed.

A is better than Z. Is it also better than B? Some writers have suggested that the place to block the slide from A to Z might be part way through the alphabet, say at G or J. Griffin proposes stopping 'at that point along the line where people's capacity to appreciate beauty, to form deep loving relationships, to accomplish something with their lives beyond just staying alive . . . all disappear'.[12] On this view, Z is a poor world not because its inhabitants' lives are imperfect, but rather because they are not *well* worth living. Put another way, their lives are insufficient.[13] A person choosing on her own behalf will accept even a crummy life, if it is marginally worth living and nonexistence is the only alternative. But she will not trade away a sufficient life, judged globally, *however* long she can extend it. Choosing on behalf of humanity, we should prefer a single good life to any number of mediocre ones. We give lexical priority to sufficient lives.

This principle is prudential. It is not a claim about distributive justice. All the parties in A, or B, or Z are *equally* well-off. We are just saying that we maximize welfare by keeping

[10] D. Dorsey, 'Headaches, Lives, and Value', *Utilitas*, 21 (2009), pp. 36-58.
[11] Huemer, 'In Defence of Repugnance', p. 914.
[12] Griffin, p. 340 n. 27.
[13] H. Frankfurt, 'Equality as a Moral Ideal', *Ethics*, 98 (1987), pp. 21-43.

lives above the sufficiency threshold. Let us call this *prudential sufficientarianism.* So long as we are choosing on behalf of humanity as a whole, this priority is defensible. We can 'adopt for society as a whole the principle of rational choice for one man' as if we were engaging in intrapersonal aggregation. It makes no moral difference that we are actually aggregating interpersonally, because whoever is born will share in the benefits, and everyone is treated the same. But following mere addition, when we move to a world with a plus, suddenly we must decide for some *at the expense of others*. That changes the name of the game.

III. MERE ADDITION AND INTERPERSONAL AGGREGATION

Let's suppose that the best letter in the alphabet, all things considered, is J. Life in J is good. Higher up the alphabet, life is better, but far fewer enjoy it. Further down the alphabet it becomes dreary, pinched, and at last hardly worth living. Assume that J is the last letter in which lives fall above the sufficiency threshold. In J+ all the original people are just as well-off as they were before, but we have added an equal number of new people, all with impoverished but nevertheless worthwhile lives. Because these lives are worth living, and they harm no one else, it seems that this must be an improvement. Yet now we face a dilemma. If we re-distribute between the groups, we will produce K. K is *not* better than J. Yet compared with J+, it 'contains greater total happiness, greater average happiness, much more equality, and a much higher standard of living for the worst-off'.[14] How could we decently refuse?

It might seem that we could refuse. Sometimes it can be worse to divide goods when this leaves too little for everyone. It is bad to distribute food or medicine evenly if this means no one receives enough to live. That amounts to *destroying* much of the good's value. The same can be true, some argue, with the goods that make for a sufficient life.[15] The J-people have lives that are barely sufficient. They are not residents of Sweden or France. They might

[14] T. Mulgan, *Future People: A Moderate Consequentialist Account of our Obligations to Future Generations* (Oxford: Clarendon Press, 2006), p. 68.
[15] Frankfurt, pp. 30-31.

live in Cuba or Kerala. The new people have lives worth living. These are not Auschwitz prisoners or famine victims. Their lives may be mediocre, but they are not hellish. They are like the lives of millions of people in today's world. Suppose they are Bangladeshis. Should we drag everyone down to the level of, say, Nicaraguans in order to raise them up?

Even if this argument against redistribution seems plausible, we can modify the mere addition paradox to make it fail. Consider

> *Interpersonal Sweatshop.* World J contains a single person, Robinson, with a life well worth living. We may add a million sweatshop workers to the population, creating World J+. All the new people will have lives worth living—just barely. They will toil ten hours every day in chilly, poorly furnished sweatshops. If we put Robinson to work as well, creating World K, we can provide all 1,000,001 with heating, padded chairs, and a radio, so that their lives, while still short of being *well* worth living, will become more pleasant.

It is not hard to believe that Robinson's life beats any number of sweatshop lives. But the new people will also have lives worth living, so adding them seems to improve the world. Once we have added them, it is hard to believe that redistribution will not also be an improvement. Think of all that suffering we could prevent, at the cost of only one person. But now our judgements contradict each other: J is better than K, J+ beats J, but K is better than J+. Either we must accept that value can be intransitive, or one of these claims has to go.

Suppose I offer you a choice:

> *Intrapersonal Sweatshop:* You may have seventy years of good life, or seventy million years toiling in a well-heated sweatshop with padded chairs and a radio.

You choose the former. Now I make a new offer: 'You can have the seventy good years, and then an *additional* seventy million years working on hard benches in a chilly sweatshop without a radio. That life isn't very good, but you will find each day barely worth living'.

You should again accept. After all, the seventy million years are, by stipulation, worth living. Suppose I make a final offer: 'Conditions in the sweatshop are tough. If you give up your seventy years of good life, I will give you heat, a padded chair and a radio'. If you are tempted by my series of offers, I have a bridge in Brooklyn I would like to show you.

It may seem that in *Interpersonal Sweatshop* it would be similarly irrational to prefer J to K, but to prefer K to J+. This is not the case. When in *Intrapersonal Sweatshop* you say that you would not exchange seventy good years for any number of years in the sweatshop, you engage in intrapersonal aggregation. All things considered, you consider this life better than that. Similarly, in *Interpersonal Sweatshop,* if we ask 'Would it be good to move from J to K?' we can judge *as if* for a single person. Everyone who ever lives will be equally well-off; our choice will not come at anyone's expense. But when we come to choosing between J+ and K, we can no longer reason in this way. If we move to K, we sacrifice the original person to benefit the multitude; if we remain in J+, we save Robinson at the majority's expense. Either way we choose, some will gain, and others will lose.

When we choose what is best for ourselves, Griffin argues, it makes sense 'to go for net gains'—to choose the best overall life. But it does not follow that we should do this when aggregating *among* individuals, or that doing so improves the outcome. When we make sacrifices on our own behalf, we pay the costs, but also reap the gains. This is not the case when we impose these costs on others. 'My life is better for enduring the hardship in order to bring off the accomplishment', Griffin explains.

> …But that I can accomplish the same if you suffer the hardship tells us nothing about your life as a whole. We need another global assessment for that. And to make it we should have to look at what the deprivation does to your life, considering your life in all its other respects too. Just by knowing

> that a certain gain and loss would be justified in my life one does not know that the gain and loss divided between our two lives would also be justified.[16]

When you choose the seventy years of good life and the unheated sweatshop, you are saying that taken together, the two are worth more to you than seventy million years in a heated sweatshop. But what would be a small benefit in your life may bulk large in mine. Suppose Robinson tells the million impoverished newcomers, 'I would rather have seventy years of good life than seventy million years working in a sweatshop. And I would rather have *both* lives as they are than *only* the sweatshop life with heat, padded chairs and a radio. That my life remains good matters more than improving yours'. The new arrivals can reply, 'That is all very well, *if* you have the seventy years of good life. But ours are the only lives we have'.[17] And they might add, echoing Mole in *The Wind in the Willows,* 'We know they are shabby, dingy little lives…but they are our own little lives, and we are fond of them!'

We accept lexical preferences within a life that we reject when aggregating among lives. 'Our fates matter equally', Griffin observes, 'that you have a chance to make something valuable out of your life matters just as much as that I do….[I]t does not matter that your life will in fact turn out less valuable than mine. In a deep sense, it is still just as important as mine'.[18] This is what is at issue in the mere addition paradox. Robinson's *life* in J is more valuable than the sweatshop lives in K. Choosing between the two—either on one's own behalf or on the behalf of possible people—it makes sense to prefer the former. But the new *people* in J+ are as important as Robinson. And these million new people will each have a strong preference for heating, padded chairs and a radio. Once they appear on the scene, their preferences, taken together, outweigh Robinson's preference for his good life.

The decision whether to move from J to K directly is like an intrapersonal decision writ large; we decide which world is better or worse for the population *as a whole*. Since the

[16] Griffin, pp. 35-36, 145-47, 180-81.
[17] R. Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), p. 33.
[18] Griffin, p. 50.

possible additional people in K do not yet exist as individuals, we can treat them, in Parfit's phrase, as 'the mere containers of an impersonal value'[19] and maximize aggregate welfare. When we choose between J and K, everyone, regardless of which one we choose, will end up with the same level of welfare. Not so when we choose between J+ and K. Now some will gain and others will lose. Such cases change the moral calculus, as Michael Otsuka and Alex Voorhoeve point out, because of the separateness of persons:

> [A] single person has a unity that renders it permissible to balance (expected) benefits and burdens against each other that might accrue to her. A group of different people, by contrast, does not possess such unity. As a consequence, some forms of balancing benefits and burdens that are permitted when these accrue to a single person are impermissible in cases where these benefits and burdens accrue to different people.[20]

Comparing J and K involves only the moral considerations that are entailed by intrapersonal aggregation, whereas the comparison between J+ and K involves interpersonal ones. Because J-lives are lexically superior to K-lives in value, we rightly regard J as better than K. But to say that a good life contributes more value to the world than a million sweatshop lives is one thing; to say it contributes more value than *helping a million sweatshop workers* quite another. To improve the J+ lives by moving on to K adds a great deal of value, but only if they have already come into existence.

Appeals to the separateness of persons are the stock-in-trade of non-consequentialists. But I believe that we can make sense of the phenomenon in desire-utilitarian terms. Moral judgments must be universalisable. What this entails, R. M. Hare maintains, is that we should accord equal weight to all preferences, adjusted for their strength. Hare recommends the Lewis method of imagining that we will occupy each party's position, and making the choice

[19] 'Acts and Outcomes: A Reply to Boonin-Vail', *Philosophy & Public Affairs*, 25 (1996), pp. 308-17, at p. 313.
[20] 'Why It Matters that Some Are Worse Off Than Others: An Argument against the Priority View', *Philosophy & Public Affairs*, 37 (2009), pp. 171-99, quoted passage at p. 179.

that we would, all in all, prefer.[21] In *Intrapersonal Sweatshop,* you can reasonably prefer a single good life to seventy million years in a heated sweatshop. Likewise, possible people could prefer the former. But in *Interpersonal Sweatshop* Robinson's preference cannot plausibly be a million times as strong as that of each sweatshop worker for heat, padded chairs and a radio. If we refused to redistribute after moving to J+, Robinson would become a *sufficiency monster.* To keep him at the higher level, we would sacrifice significant improvements to the lives of a million people. It is hard to believe that this would maximize the satisfaction of preferences, or could be best for the world.

IV. PARFIT'S DISUNITING METAPHYSICS AND THE SEPARATENESS OF PERSONS

The foregoing argument holds up only if human lives can to some extent be valued as wholes. Parfit argues that human lives are less integrated than we usually think.[22] Our sources of pleasure, desires and objective interests (if there are such things) are not all the same as they were twenty years ago, or as they will be twenty years hence. Some still connect us to our past or future selves. But other connections weaken or dissolve as time goes by. If even single lives are divided by time, then the separateness of persons seems less important. If we have a greater stake in our present well-being than in our past or future, it becomes harder for a gain at one point to compensate for a loss at another. The overall trajectory of a life becomes less important. It may then be less important whether one person bears a burden and another person reaps a benefit. Suppose that this is true. Could the distinction between types of aggregation still explain the mere addition paradox?

It might seem that my argument does not hinge on compensation occurring more easily across time than across persons. When I choose the seventy years of good life in *Intrapersonal Sweatshop,* I engage in intrapersonal aggregation. But here, we might say, I am

[21] R. M. Hare, *Moral Thinking: Its Levels, Method and Point* (Oxford: Clarendon Press, 1981).
[22] *Reasons and Persons*, part III.

not trading off gains at one point in my life against losses at another. I am trading *quantity* of life for a better *quality*. Unfortunately, it is not so simple. At the beginning, choosing the seventy good years seems like a favourable bargain. At the end, when I am on my deathbed, it will seem less attractive. Unless my dying self will have at least *some* compensation from my earlier benefits, from the trajectory of my life as a whole, won't I have aggregated at his expense? What would make this better than giving Robinson seventy years of good life at the expense of leaving a million sweatshop workers without heat and a radio?

As a first line of defence, let's assume that intertemporal compensation is altogether impossible. We can only enjoy, as Parfit describes it, 'simultaneous compensation, as when the pain of exposing my face to a freezing wind is fully compensated by the sight of the sublime view from the mountain I have climbed'.[23] The morally relevant units must then be persons at a particular time—'person slices', as David Brink puts it.[24] If I choose the seventy good years, the person slice at 9:00 AM on 2 March 2027 will benefit from my decision, as will the person-slices on 2 February 2056 and 17 October 2082, and all those in between. Even the final person-slice, moments before my body expires, may be glad to be in a well-appointed hospital, surrounded by family members and friends, rather than toiling in a sweatshop. Since he derives no pleasure from the benefits to my other person-slices, he should also feel no distress about my impending demise. So if only simultaneous compensation is possible, my choice of the seventy good years benefits *all* my person-slices from beginning to end, at no one's expense.

But surely the more plausible view is that some intrapersonal compensation is possible, even if it is harder than we usually think. Compensation within lives should still be easier than compensation between them. Though there is *some* distinction between our past, present, and future selves, these selves are far more tightly connected than they are to other

[23] *Reasons and Persons*, p. 343.
[24] D. O. Brink, 'Rational Egoism and the Separateness of Persons', in J. Dancy (ed), *Reading Parfit* (Oxford: Blackwell, 1997), pp. 96-134..

persons. Assume that I have a stake in the well-being of my past and future selves. As I lie on my deathbed, I know that my near-future self will suffer a loss. Moreover, my death will thwart many of my present desires. I thus have *some* reason to regret having passed up near-immortality in the sweatshop. Nevertheless, I can look back on a long and rewarding life. Benefits to my past self may provide some consolation and compensation. Perhaps more importantly, some of my desires concern the nature of my life as a whole.[25] As a whole, my life has gone far better than it would have in the sweatshop. Even my seventy-year-old self, while feeling some regret at his impending demise, may gain from my choice.

In short, when I choose the seventy good years in *Intrapersonal Sweatshop*, most of my future selves *unambiguously* come out ahead, and I may benefit even at the end of my life. Nothing of the sort is true in the interpersonal case. If we refuse to put Robinson to work in *Interpersonal Sweatshop*, a million workers will lose their chance for heating, padded chairs and a radio. These workers don't even know of Robinson's existence, and benefits to him do not compensate them *at all*. Even if we accept Parfit's 'disuniting metaphysics'[26]—as to some extent, I think we should—it makes a clear moral difference whether we are choosing directly between J and K, or stop off at J+ en route. Indeed, Parfit might now agree. 'As Rawls points out', he notes in his most recent work, 'if we imagine that we shall be in the positions of all the people whom our acts might affect, we shall be led to ignore the fact that, in the real world, our acts would affect different people. One person's burdens cannot be compensated by benefits to other people'.[27] That is precisely why J+ is worse than K.

V. J+ IS WORSE THAN J

If J is better than K, and K is better than J+, then J must be better than J+. Yet mere addition harms no one, and adds lives worth living. It results in inequality, but it is hard to see

[25] Brink, p. 112; D. Jeske, 'Persons, Compensation, and Utilitarianism', *Philosophical Review*, 102 (1993), pp. 541-75; B. Schultz, 'Persons, Selves and Utilitarianism', *Ethics*, 96 (1986), pp. 721-45, at p. 733.
[26] J. Broome, *Weighing Goods: Equality, Uncertainty and Time* (Oxford: Basil Blackwell, 1991), p. 43.
[27] D. Parfit, *On What Matters* (OUP, 2011), vol. 1, pp. 329-30.

why this should be bad if it leaves no one worse off. Larry Temkin has proposed that the inequality following mere addition may not be bad in comparison with A, but only in comparison with B. Since he also thinks that A may be better than B, his argument implies that betterness can be intransitive: A<A+<B<A.[28] Temkin is prepared to bite that bullet. Most philosophers have resisted the claim. Can we resist it as well?

John Broome maintains that betterness is transitive by definition. If it seems otherwise, we have failed to differentiate the outcomes.[29] Broome gives the following example:

> Maurice prefers visiting Rome to mountaineering in the Alps, and he prefers staying at home to visiting Rome. However, he does not prefer staying at home to mountaineering; if he had a choice between those two alternatives, he would take the mountaineering trips....Maurice's claim to rationality is this. Mountaineering frightens him, so he prefers visiting Rome. Sightseeing bores him, so he prefers staying at home. But to stay at home when he could have gone mountaineering would, he believes, be cowardly. That is why, if he had the choice between staying at home and going mountaineering, he would choose to go mountaineering. (To visit Rome when he could have gone mountaineering seems to him cultured rather than cowardly.)

Maurice seems to have intransitive preferences: Mountaineering<Rome<Staying home<Mountaineering. This appearance dissolves, Broome shows, once we recognize that each outcome's value depends on its alternatives. The outcomes from which Maurice thinks he is choosing, fully individuated, are

(S) Sightseeing in Rome

(M) Mountaineering

($H_{1)}$ Staying home

[28] Larry Temkin, 'Intransitivity and the Mere Addition Paradox', *Philosophy & Public Affairs,* 16 (1987), pp. 138-87.
[29] J. Broome, 'Reply to Vallentyne', *Philosophy and Phenomenological Research*, 78 (2009), pp.747-52.

(H_2) Staying home *and being a coward*

If Maurice is choosing between Rome and staying home, he is choosing between S and H_1. But if the choice is between home and the Alps, then he is choosing between M and H_2. 'If Maurice is correct in his opinion about cowardice', Broome concludes, 'it really does make a difference to the value of staying at home whether or not he has rejected a mountaineering trip. If he has, this is a cowardly thing to do, but not otherwise'. His preferences are not really intransitive; it is a matter of 'individuating the outcomes'.

Broome thinks that such cases present no real threat to transitivity. While he acknowledges that the value of an option can depend on which alternatives are available, provided that *all* options are on the table at the same time, he seems to consider it a mere matter of 'sorting them out'. [30] How should Maurice sort his choices out? Even if he individuates the outcomes, whichever option he chooses, another one will always be preferable. His problem is that a single choice—staying at home—assumes two different values in pairwise comparisons with the other choices—and yet he must decide among all three options at once. Maurice's preferences, however, are not truly intransitive. One apparent option—staying at home *without* being a coward—is illusory. Precisely *because* of one of his other alternatives—mountaineering—it is not available to him. In reality, he has three choices: H_2 (staying at home and being a coward), M (mountaineering), and S (sightseeing). Since he prefers M to H_2, and S to M, he should go to Rome.

Can we also dispel intransitivity from the mere addition paradox by individuating the outcomes? It may appear that we can by individuating K. When we compare J and K directly in *Interpersonal Sweatshop*, we are comparing one good life with K_1 (1,000,001 sweatshop lives with heating, padded chairs and a radio). But when we compare J+ and K, we are comparing one good life and a million sweatshop lives with K_2 (impoverishing one good life

[30] Broome, *Weighing Goods*, pp. 61, 100-6.

and *giving a million sweatshop workers heating, padded chairs and a radio*). The paradox arises because K takes on different values depending on the nature of the comparison. So far, so good. But this doesn't seem to rescue us from the danger of intransitive preferences. In Maurice's case, once we individuate the outcomes in H, we see that H_1 is preferable to M and S, but that H_2 is not. Since M renders H_1 unavailable, this dispels the threat of intransitivity. Here, however, *both* K_1 and K_2 are inferior to J. Since we arrive at K_2 by a series of steps each of which seems an improvement, individuating K won't solve the problem.

But now let us individuate J+ instead. Maybe, when struggling with the mere addition paradox, we are really comparing four hypothetical options rather than three:

(J) One good life

($J+_1$) One good life and a million sweatshop lives, without a property that makes it better to move on to K

($J+_2$) One good life and a million sweatshop lives, with a property that makes it better to move on to K

(K) 1,000,001 sweatshop lives with heating, padded chairs, and a radio

If $J+_1$ were available, it would be the best choice. But the inequality in J+ makes $J+_1$ impossible. While $J+_1$ would be better than J, $J+_2$ is not. Like $J+_1$, $J+_2$ contains greater utility than J, without making anyone worse off. But unlike $J+_1$, it has the drawback of making it better to choose K. That undesirable property renders $J+_2$ worse than J.

Can something can be bad simply because it makes it better to do something worse? Take the example of promises. Suppose that I, a strict act consequentialist, have a small coin in my pocket which I plan to put in an Oxfam donation box. For some reason, moved by a once-in-a-lifetime whim, I promise my daughter a stick of chewing gum instead. My daughter's face lights up, making me smile. Before the promise, it would have been best to give the coin to Oxfam. Now, the effects of breaking the promise would be such that it is, all

things considered, best to buy the gum. Impartial act consequentialists must conclude that my promise has had a bad effect. But the badness doesn't reside in the immediate state of affairs it has brought about, with both me and my daughter smiling. Rather, it consists in the promise having made it better for me to do something worse.

Now, one might object that the promise *reduces* the best alternative's expected utility, whereas moving to $J+_2$ has the effect of making one alternative—K—*better*. But that would be a mistake. While K is better than $J+_2$, it is not better than J. Just as the promise leads to a worse outcome than if I had not made it, choosing $J+_2$ leads to a worse state of affairs than if I had remained at J. Both the promise and $J+_2$ have a property that leads to an inferior outcome. That is all we must establish to say that $J+_2$ is worse than J.

That also explains why there would be nothing wrong about creating the sweatshop workers if one were unable to help them. Just as there would be nothing bad about Maurice staying home if mountaineering were not an option, there seems nothing *ipso facto* objectionable about creating a large number of people with worthwhile lives.[31] In each case, eliminating one choice affects the value of another one. In Maurice's case, eliminating mountaineering converts *staying at home and being a coward* into simply *staying home.* In *Interpersonal Sweatshop,* eliminating K affects J+'s value in a different fashion. J+'s inequality still renders a judgement in favour of the good life untenable. We are still creating $J+_2$, a state of affairs in which it *would* be better to move on to K. But a state of affairs where we *would* move on to K is not worse than J if we can no longer actually do so.

So long as the choice of K is available, then J+ *is* worse than J, because of its inequality. This is not because inequality is bad in itself. Rather, it is because inequality changes the aggregation of preferences, leading us to K. Compared with other morally relevant properties such as 'causing horrendous pain', 'changing the aggregation of

[31] D. Boonin-Vail, 'Don't Stop Thinking About Tomorrow: Two Paradoxes About Duties to Future Generations', *Philosophy & Public Affairs,* 25 (1996), pp. 267-307, at pp. 290-93.

preferences' seems a curiously abstract drawback. But it has a real harmful consequence: choosing J+ leads to a worse outcome than if we stay at J.

VI. FITTING THE PRINCIPLES TO THE POPULATIONS

The mere addition paradox has proven so troubling for moral philosophy because nearly every ethical theory has trouble accounting for all our intuitive convictions about it.[32] But no plausible principle can tell us how to maximize value when our choices will affect everyone the same way, how to maximize value when our choices will affect different people in the same population differently, *and* tell us to do the same thing in both cases.

If you are ever entrusted with the destinies of 10 billion people, all very well off, you should keep them above the sufficiency threshold. This is not a matter of distributive justice, but of maximizing aggregate welfare. You are free to add additional people and redistribute, if it improves things, but you should stop—at the very latest— before redistribution brings the quality of life below that point. Resist the siren song of those who tell you that J+ is better, because the new people will also have lives worth living. You know that once you have added them, it will become better to move on to K. Since you are still in J, and J is better than K, it is here you should draw the line. Similarly, if you could have more children, but this would mean either privileging your firstborn or neglecting them all in equal measure, you ought to stop with the ones you have. People who plan their families at all ordinarily do.

But in most circumstances, sufficiency, as Paula Casal says, is not enough; the real world is not like J, where everyone enjoys the same level of welfare. Even if you are a planner who starts off with a homogenously well-off society, it won't last: some people with poor lives will be born whatever you do. Simply maximizing the number of sufficient lives—or even the proportion of them—would accord too little weight to their needs.[33] If you have the bad luck to have already reached J+, you should redistribute and move to K. But you will

[32] Gustaf Arrhenius, 'Future Generations: A Challenge for Moral Theories' (PhD dissertation, Uppsala University, 2000), p. 4.
[33] P. Casal, 'Why Sufficiency Is Not Enough', *Ethics*, 117 (2007), pp. 296-326.

still recognize that lives in J were lexically better, and take steps to reduce the birth rate, so that as many in the next generation as possible fall above sufficiency without neglecting the worse-off. We might call this *priority-constrained sufficientarianism*.[34]

Confronted with the mere addition paradox, we should be prudential sufficientarians in the equal worlds, but prioritarians in the unequal ones.[35] This blocks the repugnant conclusion and solves the mere addition paradox. Moreover, whether our decisions affect people's identities is irrelevant. My argument thus meets the conditions Parfit set for a successful population theory, what he called 'Theory X'.[36]

Worthwhile lives are good. Lives *well* worth living are best. More worthwhile lives are better than fewer—if it increases welfare and is fair to the worse-off. Some will reject this out of hand: 'The United Kingdom is not a happier place than New Zealand', scoffs Don Locke, 'just because it contains more people'.[37] But when we are speaking of whole lives, Huemer's claim that 'more of a good thing is better' is compelling. Moreover, it provides the strongest possible grounds for the belief that extinction would be a terrible thing.

This is not just an academic matter. A few prominent scholars have suggested that runaway climate change could conceivably wipe out humanity. How seriously we should take such unlikely but catastrophic threats—and how much we should spend to avert them—depends significantly on whether our ongoing existence—and that of other species—matters.[38] At the end of *Reasons and Persons,* Parfit claimed that extinction would be far worse than a war that killed *nearly* all of us. This paper has shown why he was right.

---

[34] Cf. Paula Casal's discussion of 'sufficiency-constrained prioritarianism' *(ibid*., p. 320).
[35] We could also be 'pluralist egalitarians'. See Parfit, 'Equality or Priority'. I believe the priority view to be the better choice, but cannot defend it here.
[36] *Reasons and Persons*, p. 443.
[37] D. Locke, 'The Parfit Population Problem', *Philosophy*, 62 (1987), pp. 131-57, at p. 141.
[38] J. Broome, 'The Most Important Thing About Climate Change', in J. Boston (ed), *Public Policy: Why Ethics Matters* (Canberra: ANU E Press, 2010), pp. 101-16; M. Weitzman, 'Fat-tailed uncertainty in the economics of climate change'. Paper presented at REEP Symposium on Fat Tails, 23 February 2011, .http://www.economics.harvard.edu/files/faculty/61_REEP2011%20fat-tail.pdf. Accessed 17 June 2011.