An Intrapersonal Addition Paradox*

Jacob M. Nebel

Forthcoming in *Ethics*

Many of us want to avoid what Parfit calls

The Repugnant Conclusion: "Compared with the existence of very many people—say, ten billion—all of whom have a very high quality of life, there must be some much larger number of people whose existence, if other things are equal, would be *better*, even though these people would have lives that are barely worth living."[1]

The repugnant conclusion is an implausible consequence of classical utilitarianism. But it is not just a problem for classical utilitarians. It follows from premises whose plausibility does not depend on any utilitarian doctrine—or, indeed, on any brand of welfarism or consequentialism. That is one lesson of Parfit's mere addition paradox and of the impossibility theorems it has inspired.[2]

*For helpful comments and discussion, I am grateful to Cian Dorr, Kara Dreher, Johann Frick, Ben Holguín, Kacper Kowalczyk, Harvey Lederman, Jacob Ross, Trevor Teitel, Daniel Viehoff, Ralph Wedgwood, Jake Zuehl, two anonymous reviewers, and audiences at the University of Southern California, the London School of Economics, and New York University. I am especially grateful to Samuel Scheffler for extremely helpful feedback at several stages throughout this project.

[1]Derek Parfit, "Overpopulation and the Quality of Life," in Peter Singer, ed., *Applied Ethics*, Oxford Readings in Philosophy (Oxford University Press, 1986), pp. 145–64, at 150, emphasis original.

[2]See Yew-Kwang Ng, "What Should We Do About Future Generations?: Impossibility of Parfit's Theory X," *Economics and Philosophy* 5 (1989): pp. 235–53; Tyler Cowen, "What Do We Learn from the Repugnant Conclusion?" *Ethics* 106 (1996): pp. 754–75; Charles Blackorby et al., "Critical Levels and the (Reverse) Repugnant Conclusion," *Journal of Economics* 67 (1998): pp. 1–15; Erik Carlson, "Mere Addition and Two Trilemmas of

Unfortunately, I have no solution to the problem of avoiding the repugnant conclusion. My aim is to make the problem even more difficult. I provide a new kind of argument for the repugnant conclusion that I believe to be more compelling than the existing arguments—in particular, than the mere addition paradox. Unlike existing arguments for the repugnant conclusion, my argument does not appeal directly to controversial comparisons between outcomes in which different numbers of people would exist, or in which some people would fare better than others. Instead, my argument appeals to principles about what is good for a person under conditions of uncertainty, and to further principles that connect the prospective good of an individual to the impartial goodness of outcomes. The argument shows that, in order to avoid the repugnant conclusion, it will not be enough to reject some plausible claims about the comparative value of populations; it may also require some radical moves in the theories of prudential value and of moral and rational choice under uncertainty.

Some people are happy to accept the repugnant conclusion. If you are one such person, then you may welcome my argument. I am not happy to accept the repugnant conclusion. Nor am I comfortable rejecting any premise of the argument. I therefore regard the argument as a paradox, rather than a proof. But even if the repugnant conclusion simply leaves you cold, you may nonetheless find interest in the argument. For it raises some puzzling questions about the value of a person's life compared to her nonexistence, and about how to make decisions under uncertainty for the sake of people whose existence might depend on what we do.

I start by reviewing a version of the mere addition paradox, before turning to develop my own argument. My argument proceeds in two stages: first, an argument, structurally analogous to the mere addition paradox, for an intrapersonal analogue of the repugnant conclusion; second, an argument from the intrapersonal analogue to Parfit's repugnant conclusion. The paradox consists of both stages of the argument together. The rest of the paper is about how the argument might be resisted. I conclude with some speculations about the repugnance of the repugnant conclusion.

Population Ethics," *Economics and Philosophy* 14 (1998): p. 283; Gustaf Arrhenius, "An Impossibility Theorem for Welfarist Axiologies," *Economics and Philosophy* 16 (2000): pp. 247–66; Philip Kitcher, "Parfit's Puzzle," *Noûs* 34 (2000): pp. 550–77. For a critical perspective, see Teruji Thomas, "Some Possibilities in Population Axiology," *Mind* (2017).

1 The Mere Addition Paradox

In this section, I lay out a version of the mere addition paradox.[3] I do not necessarily endorse the argument. But having it on the table will make it easier to follow, and to appreciate the significance of, my analogous intrapersonal argument.

Consider the outcomes in table 1: *A*, *A*+, and *Z*. The rows represent these outcomes' distributions of well-being. The first two columns represent the welfare of different groups of people in these outcomes; the next two columns display the total and average well-being in each outcome. The number in each cell (if there is one) represents the welfare of the relevant group (column) in that outcome (row). (An empty cell represents nonexistence.) These numbers are supposed to be values on an interpersonal ratio scale of well-being. This means, for example, that one person's life at level 2 is twice as good as any other person's life at level 1. I assume that a life at level 100 is very good, that a life at level 2 is barely worth living (for brevity, *mediocre*), and that a life at any positive level is worth living. If you have doubts about ratio-scale measurement of well-being, please set them aside; my own argument will not require numerical representations.

Table 1: The Mere Addition Paradox

	10 billion people	9.99 trillion people	Total	Average
A	100		1 trillion	100
A+	111	1	11.1 trillion	1.11
Z	2	2	20 trillion	2

In *A*, there are ten billion people, all with very happy lives. In *A*+, those same people are better off, but there is a much larger group of 9.99 trillion people, in some distant corner of the universe, with mediocre lives.[4] In *Z*, both groups of people exist, all with mediocre lives,

[3] This version is closer to the versions in Thomas Schwartz, "Welfare Judgments and Future Generations," *Theory and Decision* 11 (1979): pp. 181–94, and Michael Huemer, "In Defence of Repugnance," *Mind* 117 (2008): pp. 899–933, than to Parfit's original. Parfit first published his version after Schwartz in "Future Generations: Further Problems," *Philosophy & Public Affairs* (1982): pp. 113–72, but his argument had already been discussed by Peter Singer, "A Utilitarian Population Principle," in Michael Bayles, ed., *Ethics and Population* (Cambridge, 1976), and Jeff McMahan, "Problems of Population Theory," *Ethics* 92 (1981): pp. 96–127, based on a draft called "Overpopulation" circulated as early as 1973.

[4] On some views, it is morally irrelevant that the same people exist in both *A* and *A*+. I do not claim that this feature *is* morally relevant. But it may make the argument—which, again, I do not necessarily endorse—more compelling to some people, and no less compelling to anyone. And its analogue in the intrapersonal case may be important.

but better than the mediocre lives in $A+$. The relevant instance of the repugnant conclusion is that Z is better than A—at least, if other things are equal. (This *ceteris paribus* clause restricts our attention to the value of well-being. If the goodness of outcomes depends on factors other than their distributions of well-being, we set them aside, by imagining the outcomes to be equally good in all other relevant respects. I omit this qualification in what follows, but it should be understood to hold throughout the paper.)

Here is an argument to the conclusion that Z is better than A. I present it in my own voice, but, again, I do not necessarily endorse it.

First, $A+$ is better than A. This is because $A+$ is better for everyone who would exist in A, and would otherwise differ from A only via the addition of lives worth living. This should make $A+$ better than A. For even if we doubt that the addition of lives worth living would, by itself, make the world better, their existence should not, intuitively, "swallow up" (as Broome puts it) the benefit to all of the A-people.[5] (So goes the argument.)

Next, Z is better than $A+$. These two outcomes contain the exact same people. The average—and therefore total—well-being is greater in Z, and this greater quantity of well-being is more equally distributed in Z, to the benefit of the (vastly more numerous) worse-off. For these reasons, Z should be better than $A+$. (So goes the argument.)

But if Z is better than $A+$, which is better than A, then Z must be better than A, by the transitivity of *better than* (which I assume throughout the paper, along with the transitivity of equal goodness).[6] Therefore, Z must be better than A. But that seems repugnant. This argument from seemingly true premises to a seemingly false conclusion is a version of the mere addition paradox.

The argument's premises are plausible, but they are far from incontrovertible. I mention some responses to the argument in section 4.1. The various responses and their problems have led debate about the mere addition paradox to somewhat of a stalemate, with nothing close to consensus as to where, or whether, the argument goes wrong.

In the next two sections, I develop a new argument for the repugnant conclusion. The first

[5] "Should We Value Population?*," *Journal of Political Philosophy* 13 (2005): pp. 399–413, at 409.

[6] The transitivity of *better than* is questioned by Stuart Rachels, "Counterexamples to the Transitivity of *Better Than*," *Australasian Journal of Philosophy* 76 (1998): pp. 71–83, and Larry S. Temkin, "Intransitivity and the Mere Addition Paradox," *Philosophy & Public Affairs* 16 (1987): pp. 138–87; "A Continuum Argument for Intransitivity," *Philosophy & Public Affairs* 25 (1996): pp. 175–210; *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning* (Oxford University Press, 2012). I defend transitivity in Jacob M. Nebel, "The Good, the Bad, and the Transitivity of *Better Than*," *Noûs* (2017).

part of the argument, developed in section 2, is an intrapersonal variation on the mere addition paradox. The intrapersonal case differs from the interpersonal case in two main ways. First, instead of comparing outcomes, we compare *uncertain prospects*. We imagine an agent whose uncertainty is distributed among multiple states of the world—mutually exclusive and jointly exhaustive propositions over which the agent has no causal influence (e.g., whether or not it will rain). Each prospect available to the agent assigns an outcome to each state of the world (e.g., the prospect of not bringing an umbrella yields an unhappy outcome if it rains, a fine outcome if it doesn't). Second, instead of asking which outcomes or prospects are better or worse, or ought to be preferred from an impartial perspective, we ask which are better or worse *for* some particular person—that is, which ought to be preferred for that person's sake by a fully rational agent concerned solely with her interests.[7] In section 3, I explain how we can derive conclusions about the impartial goodness of outcomes from claims about the prudential value of prospects.

One attractive feature of my argument is that it will not imply either premise of the mere addition paradox. So even if you reject one or both premises of the argument above, you might nonetheless find my argument compelling.

2 The Intrapersonal Argument

In this section, I first state an intrapersonal analogue of the repugnant conclusion, and then present an argument for that intrapersonal conclusion.

2.1 The Intrapersonal Analogue of the Repugnant Conclusion

I introduce the intrapersonal analogue of the repugnant conclusion with a case.

Suppose that some couple wants to conceive a child by injecting a single sperm into a single egg. Suppose that only one person could possibly originate from this pair of gametes—call her *Sally*. If they inject the sperm as planned (prospect $\mathcal{Z}$—script letters denote prospects), Sally's life will certainly be mediocre. But the couple has another, risky option (prospect $\mathcal{A}$). They can co-inject, along with the sperm, some other material that would either (in state

[7] For this gloss on prudential value, see (e.g.) Alex Voorhoeve and Marc Fleurbaey, "Priority or Equality for Possible People?" *Ethics* 126 (2016): pp. 929–54.

1) greatly increase Sally's quality of life or (in state 2) prevent the sperm from fertilizing the egg.[8] The couple is rationally confident to degree p that state 1 obtains.

The couple's options are depicted in table 2. The columns represent states of the world. The rows represent the prospects available to the couple. a is the welfare level of some very happy life. z is the welfare level of some mediocre life, which I will imagine to be "painless but drab," containing only simple pleasures like "muzak and potatoes."[9] We need not assume that these welfare levels can be represented by numbers.

Table 2: Intrapersonal Analogue of the Repugnant Conclusion

	State 1 (p)	State 2 ($1-p$)
$\mathcal{A}$	a	
$\mathcal{Z}$	z	z

Which of these prospects is better for Sally? The answer seems to depend on the value of p. If $p = 1$, $\mathcal{A}$ would of course be better for Sally, because it would guarantee her a much better life. What I want to know is this: is there some low value of p for which $\mathcal{Z}$ is better for Sally than $\mathcal{A}$? If p is low enough, should we hope for Sally's sake that the couple ensures that Sally exists, even though her life would be mediocre? On one view, the answer must be Yes. More generally, according to

The Intrapersonal Analogue of the Repugnant Conclusion: For any person S, there is some probability p such that any prospect in which S would have a wonderful life with probability p or less, and would otherwise never exist, is worse for S than a certainly mediocre life.[10]

The first thing to notice about this claim is that it's *not* repugnant. (I sometimes call it "the intrapersonal repugnant conclusion," but this should not be understood to imply that the

[8] Although the case involves an obviously unrealistic degree of idealization, the general idea of intracytoplasmic co-injection of sperm with other material is not science fiction. See, e.g., Hong Ma et al., "Correction of a pathogenic gene mutation in human embryos," *Nature* 548 (2017): pp. 413–9.

[9] Parfit, "Overpopulation", pp. 145–64, at 148.

[10] David McCarthy, Kalle Mikkola, and Teruji Thomas, "Utilitarianism With and Without Expected Utility," MPRA (Munich Personal Research Papers in Economics Archive) Paper No. 79315 (2016): 2.6, formulate a similar claim that is, under the conditions of their variable-population aggregation theorem, equivalent to the repugnant conclusion. Their conditions are in some ways similar to the principles of section 3, but rule out various kinds of egalitarianism and other departures from utilitarianism.

conclusion is repugnant; it merely abbreviates the longer name.) It is a very weak claim, which seems to me neither obviously true nor obviously false.[11]

The intrapersonal analogue of the repugnant conclusion will strike some readers as counter-intuitive, given certain background commitments. It is, in particular, hard to square with the common view that it cannot be worse for a person never to have existed. On that view, no possible outcome of $\mathcal{Z}$ would be better for Sally than any possible outcome of $\mathcal{A}$, and one possible outcome of $\mathcal{A}$ would be better for her than every possible outcome of $\mathcal{Z}$. That makes it hard to see how $\mathcal{Z}$ could be better for Sally than $\mathcal{A}$. For we would expect a prospect that is better for Sally to offer her some probability of a better outcome. But, although this difficulty may make the intrapersonal analogue theoretically suspect, it does not amount to repugnance.

Why is there such a stark difference in repugnance between the repugnant conclusion and its intrapersonal analogue? That is a difficult and important question, to which I return at the end of the paper. For now, I just want to get the intrapersonal analogue on the table, before I present a mere-addition-style argument for it in section 2.2. The intrapersonal analogue is important not because it is independently implausible, but because it leads to the (truly) repugnant conclusion—or so I argue in section 3.

2.2 Argument for the Intrapersonal Analogue

I now present an argument for the intrapersonal analogue of the repugnant conclusion.

Suppose that the couple has a third option, in addition to $\mathcal{A}$ and $\mathcal{Z}$. They can co-inject, along with the sperm, some other material that would guarantee Sally's existence. But this prospect ($\mathcal{A}$+) would have different effects on Sally's well-being depending on which state obtains. If state 1 obtains, $\mathcal{A}$+ would make Sally's life wonderful (level a+)—considerably better than the life she might have in $\mathcal{A}$. If state 2 obtains, $\mathcal{A}$+ would make her life mediocre (level z–)—considerably worse than the life she would have in $\mathcal{Z}$, but still worth living. The couple's options are depicted in table 3.

Like the mere addition paradox, the argument has two steps.

[11] Torbjörn Tännsjö, "Why We Ought to Accept the Repugnant Conclusion," *Utilitas* (2002): pp. 339–59, at 343–44, and M. A. Roberts, "Person-Based Consequentialism And The Procreation Obligation," *The Repugnant Conclusion* (2004): pp. 99–128, at 110–11, would appear to accept it.

Table 3: The Intrapersonal Argument

	State 1 (p)	State 2 ($1-p$)
$\mathcal{A}$	a	
$\mathcal{A}+$	$a+$	$z-$
$\mathcal{Z}$	z	z

First, $\mathcal{A}+$ seems better for Sally than $\mathcal{A}$, for any (nonzero) value of p. More generally, according to

The Probable Addition Principle: For any prospects $\mathcal{X}$ and $\mathcal{Y}$, and any person S who might exist in those prospects: if, in every state of the world in which S would exist in $\mathcal{Y}$, S would be better off in $\mathcal{X}$, and if, in every other state of the world, S's life would be worth living in $\mathcal{X}$, then $\mathcal{X}$ is better for S than $\mathcal{Y}$.

The judgment that $\mathcal{A}+$ is better for Sally than $\mathcal{A}$ can be supported by the following argument. $\mathcal{A}+$ would be better for Sally in one state of the world. And there is no state in which $\mathcal{A}+$ would be worse for her. For it cannot be *worse* for a person to exist with a life worth living than never to have existed. But if a prospect yields a better outcome for a person in some state of the world, and a worse outcome for her in no state of the world, then that prospect must be better for her. So $\mathcal{A}+$ must be better for Sally than $\mathcal{A}$. I reject this argument on page 20, but I hope it provides some *prima facie* motivation for the principle. I give another argument for it in section 7.[12]

Second, $\mathcal{Z}$ seems better for Sally than $\mathcal{A}+$, for some (very small) p. Suppose, for example, that p is one-in-a-googolplex. And recall that $z-$, although worth living, is considerably worse than z. Sally's life at $z-$ might, for example, contain a non-negligible amount of discomfort sprinkled throughout her otherwise painless but drab life. Under such conditions, it would be unreasonably reckless for Sally's parents to choose $\mathcal{A}+$ rather than $\mathcal{Z}$. A one-in-a-googolplex chance of a wonderful life is simply not worth nearly certain pain. More generally, according to

[12] A more direct argument might appeal to the intuition of M. A. Roberts, "The Better Chance Puzzle and the Value of Existence: A Defense of Person-Based Consequentialism," unpublished manuscript, The College of New Jersey (2018), that a greater probability of existence can, in some sense, make things better for someone even if the outcome in which she exists is not better for her. I do not, however, share Roberts's intuition.

Minimal Prudence: For any individual S and very high welfare level x, there are some mediocre welfare levels y and $y-$ (where $y > y-$) and some probability p, such that some prospect in which S is certain to exist at level y is better for S than any prospect in which S might, with any probability less than or equal to p, exist at level x, and would otherwise exist at level $y-$.

This principle is a bit of a mouthful, but only because it is so weak. It says that no matter how good some life would be, there must be *some* probability—which can be arbitrarily small—and *some* pair of mediocre lives—one of which may be considerably better than the other—such that a sure-thing of the better mediocre life would be better than a gamble that might, with arbitrarily small probability, yield the very good life but would, with near certainty, yield the worse mediocre life. This seems to me beyond serious doubt.

By the probable addition principle, $\mathcal{A}+$ is better for Sally than $\mathcal{A}$, for any p. By minimal prudence, $\mathcal{Z}$ is better for her than $\mathcal{A}+$, for some p. And betterness for Sally is transitive. So $\mathcal{Z}$ must be better for Sally than $\mathcal{A}$, for some p. More generally, the probable addition principle and minimal prudence together imply the intrapersonal repugnant conclusion, given the transitivity of betterness for a person. Call this *the probable addition argument*.

This argument is not particularly paradoxical—at least, by itself. Its conclusion, as I have emphasized, is not repugnant. I fear, however, that we must accept the repugnant conclusion if we accept its intrapersonal analogue. I justify that fear in section 3. After that, I return to the probable addition argument, focusing mostly on the probable addition principle, which I take to be the least plausible premise in the argument.

3 From Intrapersonal to Interpersonal Repugnance

In this section, I explain how the repugnant conclusion can be derived from its intrapersonal analogue. The strategy, very roughly, is to consider choices between prospects like Sally's $\mathcal{Z}$ and $\mathcal{A}$, but involving many people. For concreteness, I focus on a highly simplified example, but I explain, after giving the argument, how it is easily generalized.

Start by assuming the intrapersonal repugnant conclusion. It will help to assume a particular instance of it. To keep our numbers small, suppose (unrealistically) that the intrapersonal repugnant conclusion is witnessed by $p = \frac{1}{3}$: that is, a one-in-three chance of existing with

an a-life is worse for a person than certainly existing with a z-life. Obviously this value of p is too large to be plausible, but its particular value is arbitrary for present purposes; it will not affect the argument.

Now consider the outcomes in table 4. In A_0, there is just a single person—Ann—whose life is wonderful. In Z, there are three people—Bob, Cat, and Dan—whose lives are mediocre.

Table 4: A Repugnant Conclusion

	Ann	Bob	Cat	Dan
A_0	a			
Z		z	z	z

We are assuming that a prospect that guarantees level z is better for each of these people than a prospect in which they each have a one-in-three chance of existing at level a. That is our instance of the intrapersonal repugnant conclusion. I claim that, on this assumption, Z must be better than A_0. More generally, given the intrapersonal repugnant conclusion, we must accept the interpersonal repugnant conclusion. The argument for this claim has four steps.

3.1 Impartiality

Consider the outcomes in table 5. In each of A_0 through A_3, a single person is at level a. But it's a different person in each outcome. Each person in A_1 through A_3 is selected from the larger population in Z.

Table 5: The Same-Number Equality Claim

	Ann	Bob	Cat	Dan
A_0	a			
A_1		a		
A_2			a	
A_3				a

I claim that all of these outcomes are equally good. More generally, according to

The Same-Number Equality Claim: Any two outcomes containing the same number of people, all at the same level of well-being, are equally good.[13]

[13] Recall (from page 4) that we are assuming other things to be equal. The principles in this section are thus

I know of no plausible population axiology that violates the same-number equality claim. It requires impartiality between different possible people, but only when everyone is equally well off. It says nothing about tradeoffs, for example, between people who already exist and people whose existence depends on what we do. The principle is compatible with a wide range of views about such tradeoffs.

How could one reject the same-number equality claim? It would be absurd to suggest that some of the outcomes in table 5 are better than others. The only alternative to their being equally good would seem to be that they are *incommensurable*. But if these outcomes were incommensurable, then improving or worsening one of them might not make it better or worse than the others. Raz calls this the "mark of incommensurability."[14] Intuitively, though, if we improved or worsened one of these outcomes by increasing or decreasing the well-being of its sole member, that *would* make it strictly better or worse than the others. For example, if A_0 were better for Ann than A_1 is for Bob, then A_0 would be better than A_1. That is hard to explain, unless the outcomes are equally good. More generally, rejecting the same-number equality claim would make it hard to explain why it would be better if better lives were lived, even if by different people.

3.2 Rationality

Consider the prospects in table 6. There are three equiprobable states of the world. Prospect $\mathcal{A}$ guarantees the same outcome, A_0, no matter what. Prospect $\mathcal{A}*$ instead rotates between each of $A_1, \ldots, A_3$ from table 5, assigning an equal probability to each.

Table 6: Stochastic Indifference for Equal Risk

	State 1 ($\frac{1}{3}$)	State 2 ($\frac{1}{3}$)	State 3 ($\frac{1}{3}$)
$\mathcal{A}$	A_0	A_0	A_0
$\mathcal{A}*$	A_1	A_2	A_3

I claim that $\mathcal{A}$ and $\mathcal{A}*$ are equally good, given that outcomes $A_0, \ldots, A_3$ are equally good. This follows from a more general principle, whose formulation involves some new terminology. Say that a prospect is *egalitarian* just in case *(i)* in all of its possible outcomes, everyone

restricted to the goodness of outcomes with respect to their distributions of well-being. They should not be taken to presuppose that welfare is the only thing that matters.

[14] "Value Incommensurability: Some Preliminaries," *Proceedings of the Aristotelian Society* 86 (1985): pp. 117–34, at 121.

who ever exists is equally well off, and *(ii)* every person who might exist in that prospect has an equal probability of existing, at the same welfare levels, in that prospect.[15] $\mathcal{A}$ and $\mathcal{A}*$ are egalitarian prospects, in that sense. According to

Stochastic Indifference for Equal Risk: For any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$, if every possible outcome of $\mathcal{X}$ and every possible outcome of $\mathcal{Y}$ are equally good, then $\mathcal{X}$ and $\mathcal{Y}$ are equally good.

Every possible outcome of $\mathcal{A}$ and every possible outcome of $\mathcal{A}*$ are equally good, by the same-number equality claim. And these prospects are egalitarian. So, by stochastic indifference for equal risk, they are equally good.

This stochastic indifference principle is, I think, a minimal condition of social rationality. Rationality requires us to be indifferent between prospects that guarantee equally good outcomes—at least, when there is no risk of unfairness.[16]

Let me summarize the argument so far. $\mathcal{A}$ and $\mathcal{A}*$ are equally good, because they are egalitarian prospects that guarantee equally good outcomes. They guarantee equally good outcomes because, in all of their possible outcomes, the same number of people would exist, all at the same level of well-being. This result is important for the following reason: since $\mathcal{A}$ and $\mathcal{A}*$ are equally good, anything better than $\mathcal{A}*$ must be better than $\mathcal{A}$. We can therefore compare any prospect with $\mathcal{A}$ by comparing it to $\mathcal{A}*$. We do that in the third step of the argument.

3.3 Benevolence

Consider the prospects in table 7. We have already seen $\mathcal{A}*$. Prospect $\mathcal{Z}$ guarantees outcome Z, in which all three people live mediocre lives.

Each person in $\mathcal{A}*$ exists with probability one-third. And recall that we have assumed the following instance of the intrapersonal repugnant conclusion: a prospect of existing at level

[15] The terminology is due to Marc Fleurbaey, "Assessing Risky Social Situations," *Journal of Political Economy* 118 (2010): pp. 649–80. My definition extends Fleurbaey's notion to variable-population cases.

[16] The restriction to egalitarian prospects makes it immune to the critique of Peter A. Diamond, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparison of Utility: Comment," *Journal of Political Economy* 75 (1967): pp. 765–6. Some might object that neither prospect is perfectly fair, on the grounds that each prospect gives some but not others a chance of existing. I do not find this view at all plausible, but it would not significantly disrupt the argument. For, on this view, $\mathcal{A}*$ should still be at least as good as $\mathcal{A}$: if it is unfair to give some but not others a chance of existing, it would seem fairer to distribute this chance among Bob, Cat, and Dan than to concentrate it all on Ann. It is sufficient for my purposes that $\mathcal{A}*$ be at least as good as $\mathcal{A}$.

Table 7: Weak Pareto for Equal Risk

	State 1 ($p = \frac{1}{3}$)			State 2 ($p = \frac{1}{3}$)			State 3 ($p = \frac{1}{3}$)		
	Bob	Cat	Dan	Bob	Cat	Dan	Bob	Cat	Dan
$\mathcal{A}*$	a				a				a
$\mathcal{Z}$	z	z	z	z	z	z	z	z	z

a with probability one-third is worse for each person than a prospect that guarantees a life at level z. On this assumption, $\mathcal{Z}$ is better than $\mathcal{A}*$ for each of Bob, Cat, and Dan—that is, for everyone who might exist in either prospect. I claim that, on this assumption, $\mathcal{Z}$ must be better than $\mathcal{A}*$. More generally:

Weak Pareto for Equal Risk: For any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$, if $\mathcal{X}$ is better than $\mathcal{Y}$ for each person who might exist in either prospect, then $\mathcal{X}$ is better than $\mathcal{Y}$.[17]

I regard this principle as a minimal condition of benevolence under uncertainty. We ought to prefer prospects that are better for everyone—at least, when there is no risk of unfairness.

Although this Pareto principle is very plausible, it is probably more controversial than the previous two principles. So I will say more in its defense, before explaining how the argument concludes in section 3.4.

I offer an inductive argument for this Pareto principle.

Take any prospects $\mathcal{X}$ and $\mathcal{Y}$ in which only a single person might exist. Suppose that $\mathcal{X}$ is better than $\mathcal{Y}$ for that person. Then, intuitively, $\mathcal{X}$ must better than $\mathcal{Y}$.[18] For if some prospect is better for a person, then one ought to prefer that prospect for that person's sake. And if one ought to prefer some prospect for the sake of the only person who might exist, then, from an impartial perspective, one ought to prefer that prospect. This is just to say that our

[17] The fixed-population version of this principle is proposed by Fleurbaey, "Assessing Risky Social Situations". The restriction to egalitarian prospects makes it immune to the argument of Marc Fleurbaey and Alex Voorhoeve, "Decide As You Would With Full Information! An Argument Against Ex Ante Pareto," in Nir Eyal et al., eds., *Inequalities in Health: Concepts, Measures, and Ethics* (Oxford University Press, 2013), pp. 113–28.

[18] Some prioritarians would reject this claim (see Wlodek Rabinowicz, "Prioritarianism for Prospects," *Utilitas* 14 (2002): pp. 2–21; Matthew Adler, *Well-Being and Fair Distribution: Beyond Cost-Benefit Analysis* (Oxford University Press, 2011); Derek Parfit, "Another Defence of the Priority View," *Utilitas* 24 (2012): pp. 399–440). But weak Pareto could be weakened even further to accommodate these prioritarians while preserving its implication that $\mathcal{Z}$ is better than $\mathcal{A}$: namely, by only recommending *riskless* egalitarian prospects that are better for everyone.

Pareto principle is true for all prospects in which only a single person might exist. This claim will serve as the base case in an inductive argument.

For the inductive step, consider any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$ in which any number n of people might exist. For any such $\mathcal{X}$ and $\mathcal{Y}$, let $\mathcal{X}'$ and $\mathcal{Y}'$ be prospects just like $\mathcal{X}$ and $\mathcal{Y}$, but in which some additional $n+1$th person might exist, in a way that preserves the prospects' perfect equality. Plausibly, *if* the fact that $\mathcal{X}$ is better than $\mathcal{Y}$ for all of the n people would be sufficient to make $\mathcal{X}$ better than $\mathcal{Y}$—more generally, if our Pareto principle holds when there are n epistemically possible people—then the fact that $\mathcal{X}'$ is better than $\mathcal{Y}'$ for all of the $n+1$ people should be sufficient to make $\mathcal{X}'$ better than $\mathcal{Y}'$—more generally, then the principle should hold when there are $n+1$ epistemically possible people. For the only difference between these pairs of prospects is the possible existence of one more person. And, by the claim of the previous paragraph, the fact that $\mathcal{X}'$ is better for her would make $\mathcal{X}'$ better if she were the only one who might exist. So, if $\mathcal{X}'$ is not better than $\mathcal{Y}'$ despite being better for everyone, this should be for some reason having to do with some relation between the $n+1$th person and the others. It would otherwise be hard to see how her possible existence would prevent $\mathcal{X}'$ from being better than $\mathcal{Y}'$. But what relation is the culprit? If there were any risk of inequality, we could blame the relational fact that some might be worse off than others, through no fault of their own. But there is no such risk. The prospects are egalitarian. I therefore find it hard to see why the principle should be true for n but not for $n+1$.

I therefore believe that, for any natural number n, if weak Pareto holds for egalitarian prospects in which each of n people might exist, then it holds for egalitarian prospects in which each of $n+1$ people might exist. And I have argued that the principle holds when $n = 1$. So, by induction, the principle holds for any $n \geq 1$. This inductive argument shows that rejecting the principle would require us to think either that it fails even when only a single person might exist, or that the difference between its true and false instances lies in the addition of only a single possible person, whose existence is certain not to generate a tradeoff between different people's interests. Neither of these thoughts seems to me very plausible.

I do not pretend that this argument is decisive. If we are convinced that the repugnant conclusion is false, but that its intrapersonal analogue is true, then rejecting our Pareto principle might be our least bad option. But the option seems to me quite bad. I return to this possibility at the end of the paper. Suppose, for now, that we accept weak Pareto for equal risk.

3.4 Repugnance

In order to derive the repugnant conclusion, we technically need one final principle. This principle concerns *riskless* prospects—i.e., prospects that guarantee the same outcome in every state—such as $\mathcal{Z}$ and $\mathcal{A}$, which guarantee Z and A respectively no matter what. According to

Certainty Equivalence: For any riskless prospects $\mathcal{X}$ and $\mathcal{Y}$, which guarantee outcomes X and Y respectively, $\mathcal{X}$ is better than $\mathcal{Y}$ just in case X is better than Y.[19]

The better of two riskless prospects is the one with the better outcome. With this indubitable principle on the table, we can now wrap up the argument to the repugnant conclusion.

If $\mathcal{Z}$ is better for each person than $\mathcal{A}*$, then $\mathcal{Z}$ must be better than $\mathcal{A}*$, by weak Pareto for equal risk. And we are assuming, as an instance of the intrapersonal repugnant conclusion, that $\mathcal{Z}$ is indeed better for each person than $\mathcal{A}*$. So $\mathcal{Z}$ must be better than $\mathcal{A}*$. And, as we saw at the end of section 3.2, anything better than $\mathcal{A}*$ must also be better than $\mathcal{A}$, by the same-number equality claim and stochastic indifference for equal risk. So $\mathcal{Z}$ must be better than $\mathcal{A}$. But $\mathcal{Z}$ guarantees outcome Z no matter what, and $\mathcal{A}$ guarantees outcome A no matter what. So, by certainty equivalence, $\mathcal{Z}$ is better than $\mathcal{A}$ just in case Z is better than A. Therefore, Z must be better than A. That is an instance of the repugnant conclusion.

That is just an instance, obtained from an unrealistic instance of the intrapersonal analogue. But the argument is easily generalized. Assume the intrapersonal analogue of the repugnant conclusion. Take any number k of wonderful lives at level a. We can show there to be some number n of mediocre lives at level z whose existence would be better. Simply let $n \geq \frac{k}{p}$, where p satisfies the intrapersonal repugnant conclusion. Then consider three prospects, again called $\mathcal{A}$, $\mathcal{Z}$, and $\mathcal{A}*$: $\mathcal{A}$ guarantees a fixed population of k wonderful lives; $\mathcal{Z}$ guarantees a fixed population of n mediocre lives; $\mathcal{A}*$ assigns an equal probability to every possible k-sized population of people, all living wonderful lives, selected from the larger population

[19] This principle may seem too obvious to be worth stating. Some readers, however, appear to deny it. One associate editor of *Ethics*, for example, insists that no conclusions about the goodness of outcomes—such as the repugnant conclusion—can be drawn from claims about the goodness of prospects. Certainty equivalence seems to me a counterexample to this claim. But even if certainty equivalence is (somehow) rejected, it would seem sufficiently repugnant to conclude that prospect $\mathcal{Z}$ is better than prospect $\mathcal{A}$.

of n people. Each person's probability of existence in $\mathcal{A}*$ is $\frac{k}{n}$.[20] And $\frac{k}{n} \leq p$, where p (by hypothesis) satisfies the intrapersonal repugnant conclusion. So $\mathcal{Z}$ is better for each person than $\mathcal{A}*$. So, by weak Pareto for equal risk, $\mathcal{Z}$ is better than $\mathcal{A}*$. And, by the same-number equality claim and stochastic indifference for equal risk, $\mathcal{A}*$ and $\mathcal{A}$ are equally good. Therefore, $\mathcal{Z}$ is better than $\mathcal{A}$. So, by certainty equivalence, the guaranteed outcome of $\mathcal{Z}$ must be better than the guaranteed outcome of $\mathcal{A}$. That seems repugnant.

We have just seen how the repugnant conclusion can be derived from its intrapersonal analogue, given minimal conditions of impartiality (the same-number equality claim), rationality (stochastic indifference for equal risk), and benevolence (weak Pareto for equal risk). And we saw, in section 2, a seemingly good argument for the intrapersonal analogue: the probable addition argument. *Now* we have a puzzle.

I myself find the intrapersonal repugnant conclusion, and the premises of the probable addition argument, far less compelling than the argument of this section. I therefore suspect that the puzzle should be resolved at the intrapersonal level. So, in the rest of the paper, I ask how the probable addition argument might be resisted.

4 The Value of Existence

I begin this section by rejecting some responses to the probable addition argument based on existing views in population ethics. I then identify what I take to be the central issue in responding to the argument: the prudential value of a person's existence.

4.1 Intrapersonal Perfectionism and the Personal Critical Level

The probable addition argument would perhaps be uninteresting if every response to the mere addition paradox could be extended, in some straightforward and plausible way, to the intrapersonal case. But that is not so. Consider two examples.

[20] *Proof:* The number of equiprobable outcomes in $\mathcal{A}*$ is the number of possible k-sized combinations selected from a group of n people: $\frac{n!}{k!(n-k)!}$. Each person exists in $\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!}$ of those outcomes. So each person's probability of existence is $\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} \div \frac{n!}{k!(n-k)!} = \frac{(n-1)!}{n!} \cdot \frac{k!(n-k)!}{(k-1)!(n-1-k+1)!} = \frac{(n-1)\cdot(n-2)\cdots 2\cdot 1}{n\cdot(n-1)\cdot(n-2)\cdots 2\cdot 1} \cdot \frac{k\cdot(k-1)\cdot(k-2)\cdots 2\cdot 1}{(k-1)\cdot(k-2)\cdots 2\cdot 1} = \frac{k}{n}$.

First, consider the view that Parfit calls *perfectionism*: "Even if some change brings a great net benefit to those who are affected, it is a change for the worse if it involves the loss of one of the best things in life."[21] This view resolves the mere addition paradox (depicted in table 1 on page 3) by denying that Z is better than $A+$, given the plausible assumption that Z contains fewer (perhaps none) of the best things in life.

Does Parfit's perfectionism have a plausible intrapersonal analogue? The analogous view would seem to be the following: even if some prospect would, in expectation, bring a great net benefit to a person, it is worse for her if it lowers her probability of enjoying the best things in life. This view responds to the probable addition argument (concerning the prospects in table 3 on page 8) by denying that $\mathcal{Z}$ is better for Sally than $\mathcal{A}+$, however improbable state 1 is. But this view seems to me absurd. It is irrational to prefer prospects that will almost certainly be worse for us, in the pursuit of arbitrarily small chances of enjoying the best things in life. And it is manifestly unreasonable to choose such prospects on behalf of others. Perfectionism is simply not a plausible view of the goodness of individual prospects.

More generally, I find it much less plausible to deny that $\mathcal{Z}$ is better for Sally than $\mathcal{A}+$ in the intrapersonal case than to deny the analogous step in the interpersonal case. We can perhaps live with an inegalitarian, nonutilitarian axiology. We cannot, I think, live with an absurdly reckless decision theory.

For a second example, consider *critical-level* views.[22] Critical-level theorists argue that there is a fixed, positive "critical level" of well-being below which a person's existence makes the world worse, even though her life is worth living. This kind of view resolves the mere addition paradox by denying that $A+$ is better than A, assuming the mediocre lives in $A+$ fall below the critical level.

Do critical-level views have a plausible analogue in the intrapersonal case? According to a *personal* critical-level view, there is some fixed, positive welfare level below which a person's existence is worse for her than her nonexistence. This view would respond to the probable addition argument by denying that $\mathcal{A}+$ must be better for Sally than $\mathcal{A}$. For if Sally's mediocre life in $\mathcal{A}+$ falls below that level, and if she'd be sufficiently likely to lead such a life, $\mathcal{A}+$ may very well be worse for Sally than $\mathcal{A}$.

[21] "Overpopulation", pp. 145–64, at 163.

[22] John Broome, *Weighing Lives* (Oxford; New York: Oxford University Press, 2004); Charles Blackorby, Walter Bossert, and David Donaldson, *Population Issues in Social Choice Theory, Welfare Economics, and Ethics* (Cambridge University Press, 2005).

The personal critical-level view is dubiously coherent. For if some life were worse than nonexistence, then in what sense would its value be *positive*? We would expect our scale of well-being to be normalized in such a way that any life worse than nonexistence is assigned a negative value. If not, then we must have some other way of defining a neutral level of well-being, which does not involve comparisons with nonexistence. A few such methods have been proposed in the literature. But all of them, to my knowledge, are motivated primarily by the alleged incoherence of comparing lives with nonexistence. And the most plausible methods along these lines do not combine easily with the idea of a personal critical level.

Consider, for example, an elegant method proposed by Blackorby, Bossert, and Donaldson.[23] They imagine good and bad lives getting shorter and shorter and suggest that, as length of life gets arbitrarily close to zero, well-being approaches the same value. The value to which these shortenings converge is that of a neutral life, and is the zero level on their scale of well-being. Lives above this zero level are worth living; lives below it are not. But now suppose we introduce a personal critical level c, and that this level is set above zero in order to avoid the intrapersonal repugnant conclusion. Consider some person whose life is at level c. We are now supposed to think that it would be worse for this person if her life were shortened to a length arbitrarily close to zero, but that it would be *better* for her if her length of life actually *were* zero. This discontinuity seems to me extremely unnatural.

We have considered the most obvious intrapersonal analogues of two views in population ethics. The intrapersonal analogue of perfectionism seems absurd. The intrapersonal analogue of the critical-level view seems incoherent. I do not claim that we must therefore reject those views in population ethics. But their proponents need some other way of rejecting the probable addition argument.

4.2 Dominance, Noncomparativism, and Pseudodominance

Of the two views just considered, the personal critical-level view seems to me closer to being on the right track. It rejects the least plausible premise in the argument for the intrapersonal repugnant conclusion: the probable addition principle (page 8). This principle says that $\mathcal{A}+$ must be better for Sally than $\mathcal{A}$ because, in every state in which she would exist in $\mathcal{A}$, she would be even better off in $\mathcal{A}+$, and in every other state, her life in $\mathcal{A}+$ would be worth living.

[23] *Population Issues*, 25.

This principle is not nearly as compelling as the other premises in the argument.

To reject the probable addition principle, we must deny that a mediocre life is better for a person than her nonexistence. For suppose that a mediocre life is better for a person than her nonexistence. This would lead quickly to the probable addition principle, by

Personal Statewise Dominance: If the outcome of one prospect is better for a person than the outcome of another prospect in every state of the world, then the one prospect is better for her than the other.

If a mediocre life is better for Sally than her nonexistence, then $\mathcal{A}$+ would be better for Sally than $\mathcal{A}$ in every state of the world. So, by personal statewise dominance, $\mathcal{A}$+ would be better for her.

It would not help to claim that a mediocre life is just as good as, but not better than, nonexistence. For some mediocre lives are better than others. If some mediocre life is just as good as nonexistence, then any slightly better—but still mediocre—life would be better than nonexistence. So we could still obtain the intrapersonal repugnant conclusion.

So, to reject the probable addition principle, we must deny that a mediocre life is better than, or even as good as, nonexistence. The personal critical-level view does this, but in a dubiously coherent way: by claiming that a mediocre life is worse than nonexistence, despite being worth living. The remaining option is to deny that such a life is even comparable to nonexistence.[24] And that is what the most influential critical-level theorists (Broome and Blackorby et al.) in fact believe. They accept

Noncomparativism: One outcome is better for a person than another outcome only if the person exists in both outcomes.

[24] Wlodek Rabinowicz, "Broome and the Intuition of Neutrality," *Philosophical Issues* 19 (2009): pp. 389–411, suggests instead that certain lives are *on a par* with nonexistence, where parity is a value relation that implies comparability but rules out the standard relations of betterness, worseness, and equal goodness (see Ruth Chang, "The Possibility of Parity," *Ethics* 112 (2002): pp. 659–88). On Rabinowicz's view, however, other lives (above some zone of parity) are better than nonexistence, and that is enough to obtain a version of the intrapersonal (and therefore interpersonal) repugnant conclusion: just replace "mediocre" with "barely better than nonexistence." For other objections to Rabinowicz's view, see John Broome, "Reply to Rabinowicz," *Philosophical Issues* 19 (2009): pp. 412–7, and Jacob M. Nebel, "Incommensurability in Population Ethics," B.Phil. Thesis, University of Oxford (2015).

The most influential argument for noncomparativism goes like this.[25] If it is better for a person to exist than never to have existed, then it would be worse for her if she never existed than if she did. But if she never existed, then there would be no *her* for whom that could have been worse. Thus, if a person does not exist in one of two outcomes, then neither outcome can be better for her than the other. Call this *the metaphysical argument*. I return to it in section 5.3.

Can noncomparativists reject the probable addition principle? It might seem that they cannot. For we might seem able to strengthen our statewise dominance principle to

Statewise Pseudodominance: If the outcome of one prospect is better for a person than the outcome of another prospect in some state of the world, and is no worse for her in any state of the world, then the one prospect is better for her than the other.

This principle does, given noncomparativism, imply the probable addition principle. For if $\mathcal{A}$+'s outcome is better for Sally than $\mathcal{A}$'s in every state in which she would exist in $\mathcal{A}$, and gives her a life worth living in every other state, then $\mathcal{A}$+'s outcome is better for her in some state and worse for her in no state. (This was the argument given on page 8.)

Noncomparativists, however, should reject statewise pseudodominance. It leads to betterness cycles. Suppose, for example, that Sally's prospects are as depicted in table 8. (For concreteness, I represent welfare levels with numbers, but this representation is inessential to the argument.)

Table 8: The Pseudodominance Cycle

	State 1 ($\frac{1}{3}$)	State 2 ($\frac{1}{3}$)	State 3 ($\frac{1}{3}$)
$\mathcal{A}$	10	5	
$\mathcal{B}$		10	5
$\mathcal{C}$	5		10

In table 8, $\mathcal{B}$'s outcome is better for Sally than $\mathcal{A}$'s in state 2 and (by noncomparativism) no worse for her in any other state. So, by statewise pseudodominance, $\mathcal{B}$ is better for her than $\mathcal{A}$. By the same reasoning, $\mathcal{C}$ is better for her than $\mathcal{B}$, and $\mathcal{A}$ is better for her than $\mathcal{C}$. That violates the acyclicity of *better for Sally than*. So statewise pseudodominance must be rejected (at least, by noncomparativists).

[25] John Broome, *Ethics Out of Economics* (Cambridge; New York: Cambridge University Press, 1999), 168.

Noncomparativists are therefore not forced to accept the probable addition principle, which is the least plausible premise in the probable addition argument. And their view is supported by the metaphysical argument, which many find compelling. So noncomparativists seem well-equipped to solve our puzzle. But it remains to be seen whether they can plausibly explain why the probable addition principle is false. I discuss this question in sections 5 and 6.

5 Noncomparativist Restrictions

Noncomparativism imposes a restriction on the prudential value of outcomes. It is not obvious what noncomparativists should say about the prudential value of prospects—in particular, about prospects like $\mathcal{A}$ and $\mathcal{A}$+.

I consider two answers: in this section, that $\mathcal{A}$ and $\mathcal{A}$+ are not even comparable for Sally; in section 6, that $\mathcal{A}$+ is worse for Sally than $\mathcal{A}$.

5.1 The Certain-Existence Restriction

The simplest extension of noncomparativism to prospects is

The Certain-Existence Restriction: One prospect is better for a person than another only if she would certainly exist in both prospects.[26]

The view can be motivated as follows. A prospect's value for a person, we might think, is just its *expected* value for her—i.e., the probability-weighted average of its outcomes' values for her. And, noncomparativists should think, outcomes in which a person does not exist have *no* value—as opposed to a value of zero—for her. So a prospect in which a person might not exist has no expected value for her. For the expectation of a variable over possible outcomes requires the variable to have a value in all of those outcomes. So prospects in which a person might not exist have no value for that person. So they cannot have greater or lesser value for her, and therefore cannot be better or worse for her, than any other prospects.

[26] This seems to be the view of Charles Blackorby, Walter Bossert, and David Donaldson, "Variable-Population Extensions of Social Aggregation Theorems," *Social Choice and Welfare* 28 (2007): pp. 567–89, at 569, who write that "individual ex-ante assessments of prospects are meaningless if the person is not alive in all possible states."

The certain-existence restriction provides a simple response to the probable addition argument: $\mathcal{A}+$ cannot be better for Sally than $\mathcal{A}$, because Sally might not exist in $\mathcal{A}$. But the certain-existence restriction seems false. Consider table 9, in which we are nearly certain that state 1 obtains, in which case Sally exists.

Table 9: Problem for the Certain-Existence Restriction

	State 1 (0.99)	State 2 (0.01)
$\mathcal{A}$	100	
$\mathcal{B}$	−100	

Our uncertainty in table 9 might be highly general—e.g., about the existence of other minds or whether one is a brain in a vat—or more specific to Sally in particular—e.g., whether her mother's nearly competed pregnancy will come to term, or whether she has developed enough to be conscious. It seems clear to me that $\mathcal{A}$ is better for Sally than $\mathcal{B}$ even in the presence of such uncertainty. The mere epistemic possibility of other minds' nonexistence, or of a nearly completed pregnancy not coming to term, should not make it impossible to promote the prospective good of our loved ones or of our future children. We ought to prefer *$\mathcal{A}$ for Sally's sake*.

This judgment can be supported by the following reasoning. Prospects that share outcomes in some states of the world should be compared by simply comparing the outcomes in which they differ. More precisely, if the outcome of one prospect is better for a person than the outcome of another prospect in some state of the world, and if those prospects assign the very same outcomes to every other state of the world, then the one prospect is better for her than the other. This is true of $\mathcal{A}$ and $\mathcal{B}$ in table 9: $\mathcal{A}$'s outcome is better for Sally than $\mathcal{B}$'s in state 1, and they share the same outcome—Sally's nonexistence—in state 2. $\mathcal{A}$ should therefore be better for Sally than $\mathcal{B}$, in violation of the certain-existence restriction.

Noncomparativists might resist the claim that $\mathcal{A}$ is better for Sally than $\mathcal{B}$. For, after all, Sally might not even exist! So there might be no such person as Sally for whom $\mathcal{A}$ could be better. I will soon, on page 25, reject the metaphysical argument for noncomparativism, on which this response seems to rest. But, in the meantime, we can dismiss the response for a different reason. We can simply stipulate that Sally does in fact exist, but that this fact is unknown to the agent. (Recall that our probabilities are just the agent's rational credences.) This stipulation makes the response unpersuasive. For it is hard to see why the mere *epistemic* possibility for some agent of Sally's nonexistence—due to the agent's uncertainty about

the existence of other minds, or of the status of some pregnancy—should make it impossible, for *metaphysical* reasons, for any prospects available to this agent to be better or worse for Sally.

I therefore reject the certain-existence restriction.

5.2 The Same-State Restriction

Noncomparativists might accommodate the intuitive judgment about table 9 by adopting a weaker restriction. According to

The Same-State Restriction: One prospect is better for a person than another only if the person would exist in the same states of the world regardless of which prospect is chosen.

The same-state restriction might be motivated as follows. Define an *event* as a subset of the set of states. Suppose that Sally exists in the same states in both of two prospects. Then there is some event conditional on which every outcome of both prospects has some value for Sally. So we need not worry about an undefined expected value if that event obtains. And, if that event does not obtain, neither prospect would have any value for Sally. Noncomparativists might think that one prospect is better for a person than another prospect just in case there is some event E such that (1) conditional on E, the one prospect has greater expected value for her than the other prospect, and (2) conditional on $\neg E$, no outcome of either prospect has any value for her. This implies the same-state restriction because, if a person exists in some state in one prospect but not in another, then there is no event that satisfies both (1) and (2).

The same-state restriction would allow noncomparativists to reject the probable addition principle, while accommodating the right result in table 9. But consider table 10.

Table 10: Problem for the Same-State Restriction

	State 1 (0.01)	State 2 (0.98)	State 3 (0.01)
$\mathcal{A}$	100	100	
$\mathcal{B}$	−100	−100	
$\mathcal{B}'$		−100	−100

In table 10, $\mathcal{A}$ seems clearly better for Sally than $\mathcal{B}'$. This violates the same-state restriction because there are some unlikely states in which Sally's existence depends on which prospect is chosen. Notice, moreover, that it would be strange to admit that $\mathcal{A}$ is better for Sally than $\mathcal{B}$ in table 9 but to deny that $\mathcal{A}$ is better for her than $\mathcal{B}'$ in 10. For, surely, if $\mathcal{A}$ is better for Sally than $\mathcal{B}$ in table 9, then $\mathcal{A}$ is better for Sally than $\mathcal{B}$ in table 10: in both cases, there is some 0.99-probability event conditional on which Sally would be much better off in $\mathcal{A}$ than in $\mathcal{B}$, and she has no probability of existing otherwise. But if $\mathcal{A}$ is better for Sally than $\mathcal{B}$ in table 10, then it would be strange to deny that $\mathcal{A}$ is better for Sally than $\mathcal{B}'$. This is because $\mathcal{B}'$ can be obtained by rearranging the outcomes of $\mathcal{B}$. These two prospects are *permutations* of each other: they assign the same outcomes to states of the same probability. And, intuitively, any two prospects that are permutations of each other are equally good for a person. Therefore, $\mathcal{B}'$ and $\mathcal{B}$ must be equally good for Sally. So, if $\mathcal{A}$ is better for Sally than $\mathcal{B}$ in table 9, then $\mathcal{A}$ must be better for her than $\mathcal{B}'$ in table 10—contrary to the same-state restriction.

5.3 The Metaphysical Argument

There are other possible noncomparativist restrictions (e.g., a *same-probability* restriction) which face similar counterexamples. But we have seen enough to notice a deeper problem for noncomparativists—at least, those noncomparativists who are motivated by the metaphysical argument.

The metaphysical argument says that an outcome X is better for a person than an alternative Y only if, were Y to obtain, Y would be worse for her than X, and that something can be worse for a person only if she exists. Presumably, if this is true for outcomes, then something similar must hold for prospects. (Otherwise, we would not expect the argument to support a restriction on the prudential value of prospects incompatible with the probable addition principle.) The similar claim for prospects would seem to be that a prospect $\mathcal{X}$ is better for a person than an alternative $\mathcal{Y}$ only if, were $\mathcal{Y}$ chosen, $\mathcal{Y}$ would—or, at least, *could*—be worse for her than $\mathcal{X}$.[27] But this claim can be ruled out by the principles we advanced against the certain-existence and same-state restrictions. Consider table 11.[28]

[27] When I say that $\mathcal{Y}$ would or could have been worse for the person, I do not necessarily mean that it would or could have yielded a worse outcome. When evaluating prospects *ex ante*, in light of the agent's evidence, we assume that a prospect can be better even if, in the actual state of the world, it yields a worse outcome. That is the sense I have in mind. (This makes the condition stated in the text compatible with various responses to the "opaque sweetening problem" of Caspar Hare, "Take the Sugar," *Analysis* 70 (2010): pp. 237–47.)

[28] A case of this kind is also considered by Teruji Thomas, "Topics in Population Ethics," D.Phil. Thesis,

Table 11: Against the Metaphysical Argument

	State 1 (0.5)	State 2 (0.5)
$\mathcal{C}$	100	
$\mathcal{D}$	−100	
$\mathcal{D}'$		−100

In table 11, $\mathcal{C}$ is better for Sally than $\mathcal{D}$. This follows from our reasoning against the certain-existence restriction: we compare prospects that share outcomes by comparing the outcomes in which they differ. $\mathcal{C}$ yields a better outcome for Sally than $\mathcal{D}$ in the only state in which they differ. And $\mathcal{D}'$ is just as good for Sally as $\mathcal{D}$, because they assign the same outcomes to states of the same probability. $\mathcal{C}$ must therefore be better for Sally than $\mathcal{D}'$.

But this conclusion—that $\mathcal{C}$ is better for Sally than $\mathcal{D}'$—is hard to square with the metaphysical argument. For suppose that $\mathcal{C}$ is chosen. Either state 1 obtains or state 2 obtains. If state 1 obtains then, had $\mathcal{D}'$ been chosen, there would have been no Sally for whom that could have been worse. If state 2 obtains then there is no Sally for whom $\mathcal{C}$ could be better. So, whichever state obtains, the metaphysical argument predicts a barrier—either Sally's actual nonexistence, or her counterfactual nonexistence—to $\mathcal{C}$'s being better for Sally than $\mathcal{D}'$. But $\mathcal{C}$ *is* better for Sally than $\mathcal{D}'$. Choosing $\mathcal{D}'$ would clearly be acting against Sally's interests. So the metaphysical argument cannot be sound.[29]

This, of course, does not show that it can be better for a person to exist than not to exist. Nor does it diagnose the error in the metaphysical argument. I claim only that the argument is unsound. The conclusion that $\mathcal{C}$ is better for Sally than $\mathcal{D}'$ seems to me more compelling than the premises of the metaphysical argument, which others have found independent reason to reject. Some argue, for example, that so long as a person *does* exist, outcomes can be better or worse for her, even if she *wouldn't* have existed had those outcomes obtained.[30] Others

University of Oxford (2016) ch. 3, although he draws a different lesson from it than I do.

[29] An anonymous reviewer finds this argument objectionably similar to unsound arguments for the moral equivalence of anonymous and nonanonymous Pareto-improvements (e.g., F. M. Kamm, *Morality, Mortality: Volume I: Death and Whom to Save from It* (New York: Oxford University Press, 2002) ch. 5). Such arguments appeal to a permutation-invariance principle for welfare distributions, much like my permutation-invariance principle for individual prospects. The principle for distributions might be rejected on the grounds that it eliminates morally relevant facts about the good of particular individuals. But the analogous principle for individual prospects is not similarly objectionable. The particular state in which an outcome occurs does not seem prudentially significant in anything like the way in which the identities of particular people might seem morally significant (e.g., due to the separateness of persons).

[30] Gustaf Arrhenius and Wlodek Rabinowicz, "The Value of Existence," in Iwao Hirose and Jonas Olson, eds.,

argue that things can be better or worse for people who never exist.[31] We need not choose between these options here.

I have argued that the metaphysical argument for noncomparativism is unsound. Of course, there may be other arguments for noncomparativism. My challenge to such noncomparativists is to (a) provide an argument for their view that is not similarly undermined by the case of table 11, while still (b) restricting the goodness of uncertain prospects in a plausible way that rules out the probable addition principle. Meanwhile, in section 6, I consider a different way of rejecting the probable addition principle.

6 The Conditional-on-Existence View

We have considered three ways in which $\mathcal{A}$+ might compare to $\mathcal{A}$. According to the probable addition principle, $\mathcal{A}$+ is better for Sally than $\mathcal{A}$. This leads to the repugnant conclusion via its intrapersonal analogue. According to the noncomparativist restrictions considered in section 5, $\mathcal{A}$+ is neither better nor worse for Sally than $\mathcal{A}$. No such extensions seem plausible. Might $\mathcal{A}$+ be *worse* for Sally than $\mathcal{A}$? We already rejected one view on which that is so: the personal critical-level view. But there is another view, which is more plausible.

6.1 Conditional Expectations and Noncomparativism

We considered, on page 23, the possibility of comparing prospects by comparing their expected values for a person *conditional* on the event (if there is one) in which the person would exist no matter which prospect is chosen. This amounts to ignoring the outcomes in which a person might not exist—hence the same-state restriction. But instead of conditionalizing on an event and then comparing the resulting expected values, we could instead

The Oxford Handbook of Value Theory (Oxford University Press, 2015).

[31] M. A. Roberts, "A New Way of Doing the Best That We Can: Person-Based Consequentialism and the Equality Problem," *Ethics* 112 (2002): pp. 315–50; Marc Fleurbaey and Alex Voorhoeve, "On the Social and Personal Value of Existence," in Iwao Hirose and Andrew Reisner, eds., *Weighing and Reasoning: Themes from the Philosophy of John Broome* (Oxford University Press, 2015). This line is perhaps easier to swallow if we distinguish two senses (or contextual resolutions) of "exists": one in which unborn people never exist; another in which, necessarily, everything necessarily exists (see Timothy Williamson, *Modal Logic as Metaphysics* (Oxford University Press, 2013)). On this view, even if you were never conceived, you would have existed in a sense that allows you to instantiate modal properties; you just wouldn't have been a person or any other concrete object.

assign a value to each prospect taken separately by conditionalizing on a person's existence *in that prospect*. Let me explain.

Voorhoeve and Fleurbaey utilize the notion of a person's expected well-being *conditional on her existence* in a prospect—for short, her conditional expected well-being.[32] Whereas a person's *un*conditional expected well-being in a prospect is the probability-weighted average of her welfare levels in each of its outcomes, her conditional expected well-being is obtained by weighting each outcome instead by its *conditional* probability on the hypothesis that she exists. Equivalently, it is her unconditional expected well-being divided by her probability of existence. Informally, we simply ignore all of a prospect's outcomes in which the person does not exist, and renormalize so that the probabilities of the remaining outcomes sum to 1. According to

The Conditional-on-Existence View: One prospect is better for a person than another prospect just in case the one offers her greater conditional expected well-being than the other.

This view conflicts with the probable addition principle. Regarding our initial puzzle (table 3 on page 8), the conditional-on-existence view implies that $\mathcal{A}+$ is worse for Sally than $\mathcal{A}$, for some values of p. Sally's conditional expected well-being in $\mathcal{A}$ is a; her conditional expected well-being in $\mathcal{A}+$ approaches $z-$ as p approaches zero. The conditional-on-existence view thus implies that $\mathcal{A}+$ may be worse for Sally than $\mathcal{A}$ without claiming, with the personal critical-level view, that a life worth living can be worse than nonexistence.

Indeed, the conditional-on-existence view meshes quite nicely with noncomparativism, which the rejection of the probable addition principle appears to require. This is noted by Harsanyi in his 1977 correspondence with Ng.[33] Harsanyi argues that if he were uncertain about his own existence from behind the veil of ignorance,

> [T]he only rational decision rule for me would be to maximize the conditional expectation of my utility function on the condition that I would in fact exist.

[32] Voorhoeve and Fleurbaey, "Priority or Equality". They put the idea to quite different purposes than ours. Indeed, they define positive welfare levels as those that are better for a person than nonexistence—which rules out, by personal statewise dominance, the view considered in this section. I reject their view, and sketch an alternative that appeals to conditional expectations in a different way, in Jacob M. Nebel, "Priority, Not Equality, for Possible People," *Ethics* 127 (2017): pp. 896–911.

[33] Published in Ng's "Some Broader Issues of Social Choice," in P. K. Pattanaik and M. Salles, eds., *Social Choice and Welfare* (1983).

> …This is so because only existing people can have real utility levels since they are the only ones able to enjoy objects with a positive utility, suffer from objects with a negative utility, and feel indifferent to objects with zero utility.

The conditional-on-existence view seems to reflect this lack of concern for the nonexistent, because it gives no weight to outcomes in which a person doesn't exist; they are simply ignored. That is as it should be, if noncomparativism is true. If a person's existence isn't better for her than her nonexistence, then making her more likely to exist should not, by itself, make the prospect better for her. Increasing the probability of a person's existence may perhaps improve her prospects by making her more likely to enjoy a better life *if* she exists. But a mere difference in her probability of existence should not affect the value of her prospective existence. So it doesn't, on the conditional-on-existence view.

The conditional-on-existence view also converges with noncomparativism's verdicts about riskless prospects. Suppose that some prospect is certain to bring Sally into existence with a life worth living, and that some alternative prospect is certain to prevent Sally from coming into existence. The conditional-on-existence view implies that neither prospect is better for Sally than the other—as noncomparativists would surely believe. It has this implication because Sally's conditional expected well-being is undefined in a prospect in which she has no probability of existence: the denominator of her conditional expected well-being is zero. Since a positive value is not greater or less than an undefined value, neither prospect is better for her than the other.

We have just seen why noncomparativists might find the conditional-on-existence view attractive. I now want to consider its implications for population ethics and raise some objections.

6.2 Asymmetric Comparativism

The conditional-on-existence view has a simple analogue in population ethics: average utilitarianism. Just as the conditional-on-existence view values a prospect for a person by dividing her unconditional expected well-being by her probability of existence, so average utilitarianism values a population by dividing its total well-being by the number of people in it.

Because of its resemblance to average utilitarianism, we can expect the conditional-on-existence view to face problems much like those for average utilitarianism. And it does. Most obviously, the view has absurd implications when a person's conditional expected well-being is negative.[34] Consider table 12.

Table 12: Two Hellish Prospects

	State 1 (0.1)	State 2 (0.9)	Conditional Expectations
$-\mathcal{A}$	−100		−100
$-\mathcal{B}$	−110	−98	−99.2

The conditional-on-existence view implies that $-\mathcal{B}$ is better for Sally than $-\mathcal{A}$. But that is absurd: $-\mathcal{B}$ offers Sally a certainly miserable existence, and $-\mathcal{A}$ offers her the possibility of being better off or not existing at all. One ought to prefer $-\mathcal{A}$ to $-\mathcal{B}$ for Sally's sake, contrary to the conditional-on-existence view.

This objection is decisive against the conditional-on-existence view as stated. But it may be reasonable to restrict the conditional-on-existence view to prospects in which a person's conditional expected well-being is not negative.[35] (Call the resulting view *the restricted conditional-on-existence view*.) This restriction may seem ad hoc, but it is perhaps justifiable if there is an asymmetry between good lives and bad lives in terms of how they compare to nonexistence. According to what we might call *asymmetric comparativism*, it cannot be better for someone to have a life worth living than never to have existed, but it is always worse for someone to have a miserable life than never to have existed. The possibility of asymmetric comparativism has been largely overlooked in the literature on the value of existence.[36] I suspect this is because the main reasons why people doubt that it can be better for a person to exist appeal to metaphysical constraints that would, if true, make it impossible for it to be *worse* for a person to exist—contrary to asymmetric comparativism. I've already expressed my skepticism about such metaphysical constraints. I think that if it cannot be better for a person to exist, this must be for more specifically ethical reasons, which might allow some lives to be worse than nonexistence.

[34]Toby Handfield, "Egalitarianism about Expected Utility," *Ethics* 128 (2018): pp. 603–11, makes a similar point against the hybrid view of Voorhoeve and Fleurbaey, although not in terms of prudential value.

[35]Compare the "restricted average utilitarianism" considered by Blackorby, Bossert, and Donaldson, *Population Issues*, 5.2.8. Voorhoeve and Fleurbaey themselves restrict their discussion to nonnegative welfare levels.

[36]Except by Gustaf Arrhenius, "Can the Person Affecting Restriction Solve the Problems in Population Ethics?" in M. A. Roberts and David Wasserman, eds., *Harming Future Persons* (Ashgate, 2009), pp. 289–314, who quickly dismisses the view for the reason stated in the next sentence.

I do not claim here that asymmetric comparativism is true. I hope to explore the possibility in other work. My claim is merely that *if* asymmetric comparativism is true, then it would not be ad hoc to restrict the conditional-on-existence view to prospects in which a person's conditional expected well-being is not negative.[37] So the objection from table 12 is not decisive, unless we can rule out asymmetric comparativism.

Other objections to average utilitarianism would likely generalize even to the restricted conditional-on-existence view. But suppose we are unpersuaded by those objections. Interestingly, a new problem arises from the interaction between the conditional-on-existence view and its implications for population axiology.

6.3 *Ex Post–Ex Ante* Inconsistency

The conditional-on-existence view not only resembles average utilitarianism; it can also serve as a foundation for a version of average utilitarianism. Suppose we accept the restricted conditional-on-existence view, as well as the principles of section 3: the same-number equality claim, stochastic indifference for equal risk, weak Pareto for equal risk, and certainty equivalence. We can then derive

The Restricted Average View: For any outcomes X and Y in which well-being is equally distributed at nonnegative welfare levels x and y respectively, X is better than Y if x is greater than y.[38]

Population size, on this view, is simply ignored.

The restricted conditional-on-existence view may now seem especially promising—at least, if asymmetric comparativism is defensible. Given the principles of section 3, it yields the restricted average view, which avoids the repugnant conclusion and perhaps the most devastating problems for average utilitarianism.

A new problem, however, emerges from this derivation of the restricted average view. Consider the prospects in table 13:

[37] I say "not negative," rather than "positive," to include cases in which one's conditional expected well-being is neutral or undefined.

[38] Let X contain k people at positive level x, and Y contain $n(\geq k)$ people at positive level y. Let $\mathcal{X}$ guarantee X, $\mathcal{Y}$ guarantee Y, and $\mathcal{X}*$ assign an equal probability to every possible k-sized population living at level x, selected from the n people in Y. If $x > y$, then $\mathcal{X}*$ is better for each person than $\mathcal{Y}$, by the conditional-on-existence view. Then apply the principles of section 3 just as in the argument of page 15.

Table 13: *Ex Post–Ex Ante* Inconsistency

	State 1 (0.25)		State 2 (0.25)		State 3 (0.25)		State 4 (0.25)		Conditional Expectations	
	i	*j*	*i*	*j*	*i*	*j*	*i*	*j*	*i*	*j*
$\mathcal{A}$	19			19	1			1	10	10
$\mathcal{B}$	20			20	2	2	2	2	8	8

In table 13, there are two possible individuals, *i* and *j*, and four equiprobable states. Both prospects are egalitarian: in every outcome, everyone who exists is equally well off, and each person has an equal probability of existing at each of the same welfare levels. All welfare levels are positive, so we can set aside problems stemming from negative welfare levels.

According to the restricted conditional-on-existence view, $\mathcal{A}$ is better than $\mathcal{B}$ for both *i* and *j*, because $\mathcal{A}$ offers each person a greater nonnegative conditional expectation of well-being. By weak Pareto for equal risk, we should therefore expect $\mathcal{A}$ to be better than $\mathcal{B}$, if the restricted conditional-on-existence view is correct.

Compare, however, the outcomes of $\mathcal{A}$ and $\mathcal{B}$, considered state-by-state. According to the restricted average view—which follows from the restricted conditional-on-existence view and the principles of section 3—$\mathcal{A}$'s outcome is worse than $\mathcal{B}$'s in every state of the world, because $\mathcal{A}$ guarantees a lower nonnegative universal level of well-being. We should therefore expect $\mathcal{A}$ *not* to be better than $\mathcal{B}$, by

Minimal Statewise Dominance: For any egalitarian prospects $\mathcal{X}$ and $\mathcal{Y}$, if $\mathcal{X}$ assigns a worse outcome than $\mathcal{Y}$ to every state of the world, then $\mathcal{X}$ is not better than $\mathcal{Y}$.

This principle is significantly weaker than the standard requirement of statewise dominance, which has been called "the most basic rationality condition under uncertainty."[39] It is hard to imagine what an adequate theory of decision under uncertainty would look like without this principle, and how such a theory could be true. It seems clearly rational—if not rationally required—to prefer a prospect that guarantees a preferable outcome, at least when there is no risk of unfairness.

[39]By Marc Fleurbaey, "Welfare economics, risk and uncertainty," *Canadian Journal of Economics/Revue canadienne d'économique* 51 (2018): pp. 5–40.

In the case of table 13, the restricted conditional-on-existence view generates an *ex post–ex ante* inconsistency: comparing the prospects' outcomes state-by-state yields one verdict; comparing the individuals' prospects person-by-person yields another. Such inconsistencies are familiar to egalitarians. But familiar egalitarian explanations do not apply to this case, because both prospects are egalitarian. We seem to lack, in this case, any familiar reason not to prefer the prospect that is better for everyone, or the prospect that guarantees a better outcome. This *ex post–ex ante* inconsistency seems hard to explain.

Proponents of the restricted conditional-on-existence view might respond in one of two ways.

They might, on the one hand, reject minimal statewise dominance and maintain that $\mathcal{A}$ is better than $\mathcal{B}$, despite guaranteeing a worse outcome. Proponents of the restricted conditional-on-existence view might try to justify this strategy on the grounds that choosing $\mathcal{B}$ is not justifiable to (because worse for) each person.[40] Such a person-centered approach, however, seems insufficient to rescue the conditional-on-existence view in this case. For if $\mathcal{A}$ is chosen, we can be certain that the only person who exists would have fared better under $\mathcal{B}$! Concern for individuals' interests therefore seems not to unequivocally support a rejection of minimal statewise dominance in this case.

What if, on the other hand, proponents of the restricted conditional-on-existence view reject weak Pareto for equal risk and deny that $\mathcal{A}$ is better than $\mathcal{B}$, despite being worse for each person? This seems to me a desperate strategy. It would, at the very least, make the conditional-on-existence view significantly less interesting for our purposes. For if we were willing to reject weak Pareto, then we would never have needed to reject the probable addition principle in the first place; we could have simply accepted the intrapersonal analogue of the repugnant conclusion while rejecting the repugnant conclusion's derivation. In other words, if proponents of the restricted conditional-on-existence view reject our Pareto principle, they will have solved our puzzle only by appealing to a completely different—and independently implausible—solution to the very same puzzle. Moreover, weak Pareto for equal risk was needed in our derivation of the restricted average view from the restricted conditional-on-existence view. If proponents of the restricted conditional-on-existence view jettison this principle, they would seem to lack any straightforward route to *rejecting* the repugnant conclusion, even if they avoid the intrapersonal route to accepting the repugnant conclusion.

[40] For thinking along these lines, see Johann Frick, "Uncertainty and Justifiability to Each Person," in Nir Eyal et al., eds., *Inequalities in Health: Concepts, Measures, and Ethics* (Oxford University Press, 2013), pp. 129–46.

The restricted conditional-on-existence view seems to me significantly less plausible than minimal statewise dominance, weak Pareto for equal risk, and the other principles of section 3. I am therefore inclined to reject the conditional-on-existence view, even when restricted to nonnegative prospects.[41]

This problem is of independent significance. For the importance of *ex post–ex ante* consistency arguably lies at the core of Harsanyi's case for utilitarianism. In order to defend his choice of average (as opposed to total) utilitarianism from behind his veil of ignorance, Harsanyi appeals to the conditional-on-existence view. But, as we have just seen, this appeal to conditional expectations would make Harsanyi's own view *ex post–ex ante* inconsistent.

7 An Argument for Probable Addition

Those of us who wish to avoid the repugnant conclusion are in a difficult position. The least plausible premise in our argument to the repugnant conclusion was the probable addition principle. We have found no plausible way of rejecting that principle. We can reject the probable addition principle only by denying that a life worth living is better for a person than nonexistence. But we have seen reason to doubt the most influential argument for noncomparativism and the most obvious ways of extending noncomparativism under uncertainty.

This does not, however, mean that the probable addition principle is true. And we might find comfort in the fact that the obvious argument for the principle, which appealed to pseudodominance on page 20, turned out to be unsound. It may therefore seem reasonable to maintain that the probable addition principle is false, even if we do not know why it is false.

Unfortunately, there is a better argument for the probable addition principle—or, rather, for a slightly weaker principle that suffices for the argument to the intrapersonal repugnant conclusion. The argument has four steps, all of which concern table 14. In this case, the welfare levels are again Sally's: a is any wonderful level, d some positive quantity of well-being, $-z$ some negative welfare level, y any positive welfare level, and p and q probabilities.

First, consider $\mathcal{A}$ and $\mathcal{A}'$, but ignore state 3—i.e., suppose that $q = 0$. Most decision theorists

[41] In the framework of McCarthy, Mikkola, and Thomas, "Utilitarianism", the conditional-on-existence view delivers a view—dubbed *veiled* average utilitarianism by Thomas, "Topics"—that divides expected total welfare by expected population size. Interestingly, this view can recommend prospects that yield worse outcomes in every state *and* prospects that are worse for everyone.

Table 14: An Argument for Probable Addition

	State 1 $(1-p)(1-q)$	State 2 $(p(1-q))$	State 3 (q)
$\mathcal{A}$	a	a	
$\mathcal{A}'$	$a+d$	$-z$	
$\mathcal{A}+$	$a+d$	$a+d$	y

would agree that, for any $a > 0$, there are *some* values of $d > 0$, $-z < 0$, and $0 < p < 1$ for which $\mathcal{A}'$ would be better for Sally than $\mathcal{A}$ (again, ignoring state 3). That is, there is *some* benefit that is worth *some* chance of a life that is, to *some* degree, not worth living. This claim would be violated only by the most radically risk-averse decision theories. So much the worse, I think, for such theories.

Second, suppose that $0 < q < 1$, and consider $\mathcal{A}$ and $\mathcal{A}'$ again. Intuitively, if $\mathcal{A}'$ was better for Sally than $\mathcal{A}$ for some d, $-z$, and p when $q = 0$, then it remains so when state 3 has positive probability. After all, the value of q does not affect the *relative* probabilities of states 1 and 2 compared to each other—i.e., the ratio of $(1-p)(1-q)$ to $p(1-q)$. And it is the ratio of those probabilities that can justify the tradeoff between a possible gain of d and a possible loss of $a + z$. Since $\mathcal{A}$ and $\mathcal{A}'$ share the very same outcomes in state 3, the probability of this state should intuitively not affect which prospect is better for Sally.

Third, consider $\mathcal{A}'$ and $\mathcal{A}+$, but ignore state 1—i.e., suppose that $p = 1$. $\mathcal{A}+$ seems better for Sally than $\mathcal{A}'$, when ignoring state 1, for any positive a, d, and y, any negative $-z$, and any q between 0 and 1. More generally, a prospect in which a person is certain to exist with a life that is certainly worth living is better for her than any prospect that, conditional on her existence, makes her life not worth living.

Fourth, suppose that $0 < p < 1$, and consider $\mathcal{A}'$ and $\mathcal{A}+$ again. This introduces the possibility that Sally already exists, no matter what we do, with a wonderful life that cannot be affected by our choice. Intuitively, if $\mathcal{A}+$ was better for Sally than $\mathcal{A}'$ when ignoring state 1, then it should remain so when state 1 is possible, for any positive a, d, and y, any negative $-z$, and any probabilities p and q between 0 and 1. Since $\mathcal{A}+$ and $\mathcal{A}'$ share the very same outcomes in state 1, the probability of this state should not affect which prospect is better for Sally. It is hard to see how, if we ought to prefer $\mathcal{A}+$ to $\mathcal{A}'$ when $p = 1$ for Sally's sake, we might be permitted not to prefer $\mathcal{A}+$ to $\mathcal{A}'$ for her sake, when these prospects only differ via the introduction of a possible state in which Sally's well-being is beyond our control.

Here is what we have shown. For some values of d, $-z$, and p, $\mathcal{A}'$ is better for Sally than $\mathcal{A}$, for any a and q. That was the conclusion of steps one and two. And, for any values of a, d, $-z$, y, p, and q, $\mathcal{A}+$ is better for Sally than $\mathcal{A}'$. That was the conclusion of steps three and four. So, for some d, $-z$, and p, $\mathcal{A}+$ is better for Sally than $\mathcal{A}$, for any a, y, and q. Since the values of $-z$ and p are irrelevant to the comparison of $\mathcal{A}$ and $\mathcal{A}+$ (the value of p does not affect Sally's well-being in either prospect, and neither prospect has any chance of giving Sally level $-z$), we can state our conclusion as follows: there is some amount of well-being such that, if, in every state in which someone would exist in $\mathcal{A}$, she'd be better off by that amount in $\mathcal{A}+$, and if, in every other state of the world, her life in $\mathcal{A}+$ would be worth living, then $\mathcal{A}+$ is better for Sally than $\mathcal{A}$. This conclusion holds regardless of the relative probabilities of those states (i.e., for any q). It is therefore enough to yield the intrapersonal repugnant conclusion, given minimal prudence: for any values of a and d, there must be some probability q and some mediocre welfare levels y and z, such that a sure-thing of z would be better than $\mathcal{A}+$. Thus, for some q, a sure-thing of z would be better than $\mathcal{A}$. That is the intrapersonal repugnant conclusion.

This argument shows that rejecting the probable addition principle is not a simple matter. It is not the kind of principle that we can reasonably reject without a compelling explanation for why it is false. In order to reject the probable addition principle, we would need to know where and why the argument just given fails.

8 Conclusion: Speculations on Repugnance

The least plausible premise in our argument to the repugnant conclusion was the probable addition principle. We have found some reason to accept this principle, and no good reason to reject it—other than the fact that, given some other highly plausible principles, it leads to the repugnant conclusion. I am unwilling to accept the repugnant conclusion. Nor am I comfortable rejecting any premise of the argument. I therefore regard the argument as a paradox, to which I have no satisfactory solution.

I want to conclude by suggesting, in light of this predicament, a possible avenue of further research.

We observed, on page 6, a stark difference in plausibility between the repugnant conclusion and its intrapersonal analogue. Why is there such a difference? Answering this question

might help us to isolate where, in our argument to the repugnant conclusion, the repugnance seeps in.

I speculate that the answer has to do with our concern for certain ingredients of well-being. We care very strongly about the existence of the things in wonderful lives—things like loving relationships, creative activities, and sophisticated pleasures. But perhaps we do not value these things—primarily, at least—because they are good for the people whose lives contain them. Perhaps we value these things primarily as *impersonal* goods. This impersonal picture might explain why, in the intrapersonal case, we are willing to deprive Sally of any chance of enjoying these things, for Sally's sake, but, in the interpersonal case, are unwilling to deprive *the world* of these things.

On this view, the feeling of repugnance seeps in with our Pareto principle. In the choice between $\mathcal{Z}$ (which offers each person a certainly mediocre life) and $\mathcal{A}*$ (which offers each person a tiny probability of existence with a wonderful life), we are unwilling to judge that $\mathcal{Z}$ is better than $\mathcal{A}*$, even if we are willing to accept that $\mathcal{Z}$ is better for everyone. For $\mathcal{A}*$ ensures that the world contains the kinds of things that fill wonderful lives, whereas $\mathcal{Z}$ deprives the world of all such things. Perhaps we care much more about the existence of such things than we do about the people who get to enjoy them.

This diagnosis of our sense of repugnance turns a common objection to classical utilitarianism on its head. The classical utilitarian is sometimes said to regard people as mere containers of value. She wants to improve people's lives not because she cares about people for their own sakes, but rather because she cares about the value—namely, well-being—contained in their lives. The repugnant conclusion seems to illustrate this objectionable feature of classical utilitarianism. The classical utilitarian wants the world to contain as much value as possible, so she wants to create as many value containers as possible. That is why she is led to the repugnant conclusion, which (according to conventional wisdom) is false and repugnant precisely because people are *not* mere containers of value.

According to the present diagnosis, however, the repugnant conclusion strikes us as repugnant precisely because people *are*—to us, in certain respects—containers of value. They are containers of goods that we value more highly than the interests of the very people whose lives they fill. This diagnosis makes me uncomfortable: I find it morally perverse to care more about the things in people's lives than about people themselves. The argument of this paper suggests a possible way in which the classical utilitarian might accept the repugnant conclusion precisely because she cares instead about each person for her own sake—in par-

ticular, about each person's interest in coming into existence, or in being made more likely to exist, with a happy life. This would provide some confirmation of a claim due to Sidgwick: "a Utilitarian …never has to sacrifice himself to an Impersonal Law, but always for some being or beings with whom he has at least some degree of fellow-feeling"[42]—sometimes, though, for beings who might never exist.

[42] *The Methods of Ethics* (London: Palgrave Macmillan UK, 1962), 500–1.