[DRAFT 2017.12.28]

Modal Ethics

Melinda A. Roberts

robertsm@tcnj.edu

Introduction

Purpose of this book. To think properly about what makes one possible world—or possible *future*, or *outcome*—morally better than another, and from there, about what we morally ought to do, we can't take into account *only* what's in our immediate vicinity—*only*, that is, what is going on in the *actual* world. We must at least glance every so often at the horizon as well. We must take into account what is going on in other possible worlds as well, including worlds so unlike our own—so, we can say, *distant* from our own—that their populations and ours are completely disjoint.

For what happens *there* can affect, indeed, *change*, what happens *here*. At least: knowing what happens *there* can affect, indeed, change, our understanding of what happens *here*. An outcome that would otherwise seem to us to be worse becomes *better*; an act that would otherwise seem to us to be wrong becomes *permissible*.

The moral significance of merely possible *people* is widely acknowledged. The phenomenon of action-at-a-*very*-great-distance as it applies to possible *people* cannot, I think, seriously be questioned. In a sense, what I am doing in this book is simply inquiring how that argument translates to merely possible *worlds*. But we should take a moment to review just how the argument works in connection with merely possible *people*.

Consider two worlds, one, the uniquely *actual* world, w1, containing a number of well-off people and the other, a *merely possible* world, w2, that makes each one of those people still better off and that contains some additional people as well. The lives in the additional cohort are unambiguously worth living. But they aren't nearly as wonderful as the lives of the people in the original cohort. Suppose,

then, that we have p1 . . . pn, the original cohort, existing in both worlds, and q1 . . . qm, the additional cohort, existing in just w2.

Graph 1: Addition Plus		
Wellbeing	w1	w2
+11		**p1…pn**
+10	**p1…pn**	
+1		**q1…qm**
	q1...qm**	

Boldface type in Graph 1 and throughout means that the indicated person does or will exist at the indicated world. *Italics* and an asterisk (*) mean that the indicated person never exists at the indicated world.

Now, we can immediately evaluate w1 as, in one respect, morally defective. The original cohort p1 . . . pn have been made less well off when they could have been made more well off. In w1, those people suffer—or we might say *incur*, if we think that *suffering* and *wellbeing* may be related but independent concepts—an *avoidable loss*, or *diminution*, in wellbeing. And one might think as well that w2 itself looks pretty stellar: it avoids the loss on behalf of the *original* cohort, and it can't, we think, be *worse* for the *additional* cohort to have lives that are unambiguously worth living than it is for them never to exist at all. If anything, w2 makes things *better* for q1 . . . qm than w1 does. (I for one support that last point. But I should note that many philosophers seem to find it an obstacle to comparison that q1…qm never exist at all in w1; they think that fact somehow makes q1 . . . qm ineligible subjects for discussion.) So it may seem we can easily conclude that w1 is worse than w2—and accordingly that the choice that produces w1 is wrong.

But that inference is far too quick. The evaluator should insist in this case that he or she does not have enough information to make any judgment at all about this case. What's missing?

Well, for all we are told, w1 and w2 *do not exhaust the alternatives*. (And even if we are *told* they *do* exhaust the alternatives, we will want—in certain cases, including some *nonidentity* cases—to challenge that stipulation. But in the case at hand the important thing is that we've been told nothing at all.)

So what if they don't? Don't we still know enough compare the two outcomes we *do* know something about?

No. For suppose there exists a third world w3—still another merely possible world, alongside w2, given that w1 is, by hypothesis, the uniquely actual world—that makes the additional cohort, q1 . . . qm, much better off without making the original cohort, p1…pn, too much worse off? Suppose, in other words, that the case, once all alternatives are actually unveiled, looks like this.

Graph 2: Addition Plus Complete			
Wellbeing	w1	w2	w3
+11		**p1…pn**	
+10	**p1…pn**		
+6			**p1…pn, q1…qm**
+1		**q1…qm**	
	q1…qm**		

w3 now before us, we can now see moral defects in w2 that we were not in a position to see before. We can now see that perhaps w2 after all *isn't* morally better than w1 and that perhaps the choice that ends in w1 *isn't* after all wrong. This is not to see that agents are *obligated* to bring the additional cohort into existence—or that their existence would make the world better. Rather, it's to say that, if agents *do* bring them into existence—e.g. at w2—then it won't do to make them *worse* off when, at no unduly great cost to anyone else, that is, to the original cohort, agents could make them *better* off instead.

Addition Plus shows that *merely possible* people have *moral status*. In doing so, Addition Plus refutes a version of moral actualism. Actually it refutes *two* versions of *moral actualism*, the one that insists only *actual* people have moral status, *and* the one that insists that for purposes of evaluating the act that results in w1, only people who *exist under that act* have moral status.[1] Of course we must

[1] Caspar Hare nicely refutes both forms of moral actualism. One principle asserts that only people who do or will exist at the uniquely actual world matter morally. This is the view

do well for actual people; they clearly have moral status. But what Addition Plus shows is that the merely possible have moral status as well.

People who will never *actually* exist, *merely possible* people residing only in *merely possible* worlds, thus command our moral attention. What is going on with them *there* can affect, indeed *change*, what we might otherwise take to be an uncontroversial evaluation of w1. Blinding ourselves to w3, we might have immediately but mistakenly concluded that w1 is worse than w2. Taking w3 into account, we may well now conclude—the door is now open to our concluding—that w1 *isn't* worse than w2 and that the act that results in w1 is *permissible* and is perhaps even *obligatory*.

Now, that last point may seem problematic. If the merely possible have moral status, and if q1 . . . qm are better off in w3 than they are in w1 or w2, *oughtn't* agents bring them into existence—*and* treat shower them with additional wellbeing? Aren't agents *obligated* to go for w3?

No. I have argued elsewhere that the point that the merely possible have moral status, though itself clearly correct, *does not mean* that the world in which they exist is morally better than the world in which they don't, or that we are under an (other things equal) obligation to bring them into existence. To say we—and they—have moral status is just to say that, other things equal, it makes the world better, or that agents ought, to add to their wellbeing stock in a case where their doing so avoids their *existing* at a lower wellbeing level—*existing and diminished wellbeing* being the key nexus here, the combination of factors that makes for moral salience.

It's the same obligation our parents had in respect of you and me: they weren't *obligated* to bring us into existence but—given that they *did* bring us into

---

that is at odds with Wlodek Rabinowicz's normative invariance. The other asserts that the people who matter morally for purposes of evaluating an act are the people who do or will exist if that act is indeed performed. Elizabeth Harman adopts such a claim as an implication of her "actual future principle."

existence—they can't treat us like chopped liver (at least, not until we are eighteen or so).

It's thus a mistake, I have argued, to divide people into two classes: to say, e.g., that *actual* people matter morally, and the *merely possible* not at all. But we still need to make a distinction to explain why q1 . . . qm's loss in w1 *doesn't* make w1 worse, but a loss of a similar magnitude and to the same individuals in w1 *does* make w2 worse (and, to take the analysis one step further, why p1 . . . pn's loss in w3 makes w3 worse). What's *not* a mistake is to divide up particular *diminutions in wellbeing*—or *losses*—between those that have moral significance and those that don't. It would have been a *loss* to me had my parents not happened to bring me into existence. But it would have been a loss I incurred as a never-existent person (a loss I incurred at a world where I never exist). It therefore (according the loss/existence nexus; under, that is, the principle I have proposed under the title *Variabilism*) would not have been a *morally significant* loss. Similarly, the loss q1 . . . qm incur in w1—concededly a *loss*—is not a *morally significant* loss. In contrast, the loss those same people incur in w2—the loss we are in a position to see *only if* we take w3 into account; *only if*, that is, that we understand that their loss in w2 is *avoidable*—is a loss they incur at a world where they exist. As such, it's a loss that has full *moral significance*, bearing *not just* on how we evaluate the merely possible w2, where those people do exist, but also on how we evaluate the actual w1, where those same people *never exist at all*. *That* is spooky action at a distance—action at a *very* great distance.

These points I have made before. But an issue has been whether these points can be fit into a *teleological* approach, that is, a theory that aims not just to evaluate certain *acts*, or *choice*, as permissible or wrong—though that would be a lot—but rather to say what makes one possible *world*, or *future*, or *outcome*, morally better than another—and from there, perhaps, to work our way back to saying what is right or wrong for at least some cases. (Or are they just more deontic, feminine or feminist, blather?) The aim of this book is to explore that question.

It's not just my pride that's at work here. For one thing, I think our telic and deontic evaluations are *very closely connected*. (As I am using and

understanding my terms, it makes no sense at all to say that the one act leads to the morally better world but it's perfectly fine to perform that act; or that the one act is obligatory but in no way makes the world in which that one act is performed morally better.) I also share with the teleologists a *basic maximizing intuition*. Moreover, while *aggregation*, or *additivity*, as a way of determining when we have done more—when we have *maximized*—is a principle that I have long been skeptical of but one that I at the same time recognize as useful (if not essential) for insuring against certain transgressions (e.g., failures of transitivity in our overall betterness-between-outcomes relations).

So there's all that. But the really critical goal for me is to deal with a potential inconsistency objection against the sorts of theories that come out where I think that they should come out in connection with—for example—*Addition Plus.*

Is it, e.g., conceptually wrongminded, indeed, inconsistent to insist that we take the *full modal landscape* into account before we can successfully work through many of the critical problems in population ethics? I argue, here, that it's not.

I make that argument specifically in reference to the *nonidentity problem* (Part I below) and the *procreative asymmetry*, and specifically the happy child half of the asymmetry, that is, the *neutrality intuition* (Part II below).

Thus, setting up the *nonidentity problem*, philosophers give us a case that includes the options of bringing no one into existence, bringing one person p into existence at a certain wellbeing level and bringing a distinct person q into existence at a still higher wellbeing level. We have the strong sense that the second option is somehow wrong and somehow worse than the third option. The most important forms of the nonidentity problem really do not stray from that model. At the same time, we may also have the intuition that bringing ever more additional happy people into existence doesn't make things better—which is itself closely related to the further intuition that the wrong act, the lesser option, must be tied to make things worse for some existing or future person—that simply leaving a person out of existence, and imposing, in that sense, some sort of loss on them, isn't enough to make a choice wrong. But that thought in mind, and going back, now, to the nonidentity problem, we now seem committed to the view that the second option

isn't wrong, and isn't lesser, at all. And we are then told that we have a paradox—a paradox we can resolve only by abandoning some deeply held intuition or another.

But that's just not so. At least I will argue in Part I that the most important forms of the nonidentity problem—those in which the second option is clearly wrong—are also those in which critical information is left out of the story. The story, that is, makes every effort to keep us from glancing up at the horizon. We are forbidden, more or less, from noticing a critical fourth option: the one in which the very same person that is left less well off under the second option is made better off. Once we notice that fourth option, we can preserve all our relevant intuitions, both the intuition that the "bad" act must be "bad for" someone—that is, the person-affecting intuition—and the intuition that the second option is wrong and is indeed worse than the third.

Now, it might seem that this way of approach the nonidentity problem involves us in a certain technical problem: how can the existence of the fourth option effect—indeed, *change*—our evaluation of how the second and the third options themselves compare? How can the third option *better* than the second if the fourth option *is* part of the case, but *not better* than the second if the fourth option *isn't* part of the case?

Help in addressing this problem, and indeed avoiding inconsistency, comes in the form of what I will call the *accessibility axiom*. That axiom will insure that the vocabulary we use in describing what is going on in this pair of cases and others is sufficiently exacting. Once we do that, consistency is trivial.

* * *

We must, accordingly, in evaluating any one world and any one act that is performed at that world keep an eye on the horizon. But another point is just as important: we can't *overweight* the significance of what it is we are looking at. It's just as much a mistake to think that something is significant when it's *not* as it is to think that something is insignificant when it *is*.

Some of what goes on *there*—in very distant possible worlds—really *doesn't bear on* what we say about what happens *here*—at the actual world or

indeed at any world we happen to be evaluating—and it would be a mistake to think that it did.

To think that saving a child's life at the expense of, e.g., the child's leg is a lesser outcome, or wrong, just because there exists *some* very distant possible world somewhere where we save the child's life *without* sacrificing the leg might thus be a mistake. Where that very distant possible world is *fantastical* or, we can just say, *inaccessible*, relative to the actual world—where, that is, no agent or collection of agents can get us from the actual world to that very distant alternate world in the absence of magic or some like suspension of the laws of nature—it seems plausible to think that the fact that that world exists exerts *no deflationary effect whatsoever* on the value of the actual world. The same point holds, not just for the actual world, but also for *whatever* world it is we happen to be evaluating, actual world or not. We don't deny the fantastical but inaccessible world exists; we don't deny, that is, that magic, e.g., is *possible*. We just insist that in that particular case the phenomenon of action-at-a-very-great-distance has lapsed. We just insist that the fact that that fantastical world is there *doesn't* mean that there's anything wrong with, or anything morally deficient in, what is here. Despite the loss of the leg, it's unconditionally *morally wonderful* to save the child's life; the world that includes that story can reasonably be ranked morally second to none across the *full modal array*; the loss of the leg alone *doesn't* entail that the world in which the leg is lost is demoted to a lower position in the great betterness ranking of *all possible worlds*, fantastical or not.

Why is this important? Consider the *procreative asymmetry*, according to which it doesn't make things better, and we aren't obligated, to bring additional *happy* people into existence, but it does make things better, and we are obligated, not to bring additional *miserable* people—people, that is, whose lives are *less* than worth living—into existence. As we shall see, it's easy enough to preserve the second half of the asymmetry—the miserable child half. Almost all standard and non-standard approaches manage to do that. What's hard is to do that while also preserving the first half of the asymmetry—the happy child half.

It's the happy child half of the asymmetry—in the refined form that Broome calls the *neutrality intuition*—that I will focus on here. It turns out that critical to making a certain variation on the neutrality intuition work—what I will call *narrow neutrality* in what follows—is a certain *inversion* in what we may think of as the value that a person's existence at a given world contributes to the value of that world itself. We shall need to say that in some cases the existence of a person whose life is clearly worth living—a happy person—will *deflate* the overall value of the world itself. That shall happen when, for example, agents **had the *real* option, the *accessible* option—*not* the *fantastical* option**—of making things still *better* for that particular person—making the child still *happier*—and have instead opted to make things *worse*. But what we shall not want to say—what we can't *plausibly* say—is that that contributory value that comes with that particular (positive) wellbeing level in the one world shall remain constant across the full modal array. In a case where there exists no such further real option, we'll need to say, instead, that a person's existence at the same wellbeing level *doesn't* have a deflationary effect on the overall value of the world. –And here we have an instance in which it's of critical importance *not* to envision *everything* on the horizon as exerting an effect at a very great distance on the here and now.

But—in order to make this position plausible—we shall **need to navigate around a certain inconsistency that** it might otherwise seem we shall run up against. How can the same person's existence at the same wellbeing level at the same world **have a deflationary effect on the value of another world in one case but not in another? Again we shall rely on the *accessibility axiom* to insure that the vocabulary we use is sufficiently exacting. Once we do that, consistency is, again, trivial.**

* * *

These are the basic themes of this book. You may come away from this introduction thinking that the book itself is going to be similarly ridiculously abstract. In fact, though, I know of no other handful of issues in moral philosophy in general, or population ethics in particular, that has the potential to create so much practical havoc. Resolutions that are now considered standard lead to results that

seem—even to their own authors—grossly counterintuitive and indeed quite bent. Really? It's wrong for a woman to have no children rather than five? Five rather than ten? Progressive liberal white male philosophers really want to sign on to a moral theory that instructs their eighteen-year old students that early abortion, indeed, even contraception, is wrong? Those same guys really want to try to convince people to make enormous sacrifices now for the purpose of preventing climate change, not for the sake of people who do or will exist and will then suffer the effects of our doing nothing, but for the sake of those who will otherwise *never exist at all*? You really want to take the position that the neighbors next door, if they were fully rational, would understand that any obligations they may have thought they had in respect of their little dog can be just as well satisfied by their simply *replacing* that little dog by another exactly similar but nonidentical little dog? (Yes, euthanasia is, by definition, painless. Yes, we can create in the lab a new little dog that's so similar to the original that, like a new pet minnow, the family will hardly know the difference. But really. We've gone off track in our moral theorizing if we don't think that this is an easy case. And, no, it's *not* enough to say that replaceability would cause the *child-owner* of the dog distress. It's what the dog, not the child, who is the central figure in this case.)

Of course we *should* sacrifice. But we should sacrifice *not* in order to avoid morally *insignificant* losses—*not* in order to bring ever more new people into existence—*but rather* in order to avoid morally *significant* losses on behalf of people—to do better, that is, rather than worse for those people (including those dogs!) *who do or will exist*.

We thus need an alternate approach, a non-standard approach, an approach that puts *theory and intuition on the same page*. The purpose of this book is to propose such an approach.

More on modal ethics. I here describe in very rough terms aspects of the approach I am calling *modal ethics* and note certain distinctions between modal ethics and other contemporary approaches that—like modal ethics—start with the basic maximizing intuition.

Modal ethics and moral actualism. Modal ethics recognizes that the plights of people who never exist at all may make a difference to the moral evaluation of choices agents make in respect of people who do or will exist. It accepts spooky action at a *very* great distance. Whether one possible future or *outcome* is better than a second is (in some cases) a matter determined in part by the availability or *accessibility* to agents of still a third outcome. (What Temkin calls the "intrinsic aspects" of the pair of outcomes under scrutiny are thus not, according to modal ethics, *sufficient* to determine the ranking of those outcomes.) The accessibility of that third outcome Z may make outcome X *better than* Y when X would otherwise have been *exactly as good as* Y.

Modal ethics thus contrasts with an *actualist* ethics, one that restricts the domain of those who matter morally to those who do or will exist at the actual world or (alternately) at the outcome at which the act under evaluation is itself performed. Modal ethics recognizes, then, the moral status of *all* possible people: they *all* matter morally and in exactly the same way.

Independence axiom. Modal ethics seems at odds with classical utilitarianism, and in critical respects it is. But there are points of agreement. For example, modal ethics accepts the basic maximizing intuition, and it agrees with classical utilitarianism that all people, including the merely possible, matter morally and in exactly the same way.

It might seem that modal ethics is at odds with classical utilitarianism regarding the sort of on-and-off, spooky action at a very great distance we have described above. It might seem that classical utilitarianism rejects the phenomenon, whereas modal ethics insists that it's real. According to classical utilitarianism, if X is exactly as good as Y when Z isn't an accessible outcome, then X is exactly as good as Y when Z is an accessible outcome. The change *there*—in what worlds are accessible—cannot effect a change *here*—a change in the overall goodness of—e.g.—the *actual* world.

The *independence axiom* may have a certain intuitive appeal that makes it hard to reject. But the clearer problem with rejecting the independence axiom emerges when we consider the *consistency* of the theory that denies it.

Of course we shall preserve consistency. But part of what will be argued in what follows is that we can do that and still recognize the very sort of action-at-a-very-great-distance that we need to recognize in order to maintain a certain version of the neutrality intuition, specifically, *narrow* neutrality. (Here, what I am calling the *accessibility axiom* comes into play.)

Contemporary consequentialism. Contemporary consequentialists, including Broome and Temkin, want their theories to have the capacity to account for values like fairness, equality and, perhaps, priority.

What they haven't seen their theories as clearly accommodating—though Temkin has made every effort, by including in his overall theory what we calls an *essentially comparative view*—intuitions in respect of our *existential* values. (We have always wanted to be *existentialists* and this is our chance.) We think, that is, that it's better to increase the stock of wellbeing of an *existing* child by rescuing that child from a burning building than it is to increase the stock of wellbeing of a *never existing* person by way of bringing that person into existence. Perhaps in both cases we increase their stock of wellbeing from nothing at all—from +0—to some perfectly reasonable amount—to, say, +10—insuring a very nice life for each. But we still want to say that saving the child from the burning building is the better world and indeed the obligatory choice. *That's* our existential intuition at work.

But contemporary consequentialists haven't clearly found a way to make sense of that intuition. Broome indicates that he has the intuition—he articulates it as the *neutrality intuition*—but he argues that it's an intuition we must in the end reject as inconsistent. Temkin thinks a correct theory must include not just an *essentially comparative* component but an *internal aspects* component as well, one that means that the evaluation of whether it's better to save the existing child from the burning building at least shall become very complicated.

My position here is that the intuition—reformulated in terms of narrow neutrality—is after all consistent, and that it's application in a given case isn't particularly complicated. There are no moral plusses to bringing the additional happy child into existence. There are plenty of moral plusses to saving the existing child from the burning building.

Additivity. Classical utilitarianism has an additive structure. I have previously contrasted that sort of theory with person-affecting, or person-based, consequentialism. In this respect, my overall view has been structurally similar to Nils Holtug. Holtug, though, adopts a "wide" person-affecting view and I adopt a "narrow" person-affecting view, a distinction that means that we will approach (e.g.) the *nonidentity problem* itself in different ways.

The importance of modal ethics for resolving the nonidentity problem cannot be understated. Any view that blinds us to what is happening on the horizon and understanding that what happens *there* affects what happens *here* is bound to fail. Two points follow. First, arguments that *promote* the nonidentity problem by averting our gaze away from the distant horizon need to be understood as deeply flawed: they are unsound and hence never compel us to accept the problematic result to the effect that a clearly wrong act is not wrong at all, or the result that the clearly worse outcome is not worse at all. Second, solutions to the problem that again disallow information regarding the distant horizon are going to fail for some forms of the problem. But it doesn't follow that solutions that don't disallow that information are going to fail as well.

Thus the two outcome, "but for" counterfactual test for determining whether a person has been made worse off, or harmed, by an act is wildly implausible. The argument that promotes the nonidentity problem by endorsing that counterfactual test, or the solution that limits itself to that counterfactual test, inherit that implausibility. To determine harm, it is not enough to look at what *would* have been had the act under scrutiny not been performed. Rather we must take into account what *could* have been: whether harm has been imposed must be tested against the full array of possible outcomes accessible to the agent or agents at the

critical time. It's silly for philosophers not to be ready to do this. Lawyers recognized this not very sophisticated modal point decades ago. Surely we can keep up with them when it comes to thinking clearly about possible worlds!

Two types of nonidentity problem. It seems we can distinguish two types of nonidentity problem. Johann Frick, Caspar Hare and Shamik Dasgupta disagree with me on the "can't-do-better" type. But I am not sure they realize how narrow that category is. I think the problem is that they haven't recognized the two types. David Boonin and David Heyd agree with me on this type. But I think that they too fail to recognize that there are two or more distinct problem types. I argue that the latter type, the "can't-expect-better" type, covers far more ground and far more clearly includes the intuition that what is done is wrong—and that in all instances of that particular type we can show harm in an intuitive, comparative, "worse off" sense of that term.

Modal Ethics

Part I: The Nonidentity Problem

Chapter 1

Intuition and Identity

1.1 *Goals, organization.* The most challenging nonidentity cases are those in which the act under evaluation is *clearly wrong* and the world in which that act is performed is *clearly worse* but we seem unable to find *any basis* on which to say *why* that's so *without* abandoning a certain deeply held intuition—the *person-affecting intuition.*

I believe that to solve the nonidentity problem but retain the person-affecting intuition we need to do two things. First, we need to recognize certain details—*modal* details—inherent in that class of most challenging cases, details about worlds *beyond* the world as it is and the world as it would otherwise have been. And, second, we need to formulate the person-affecting intuition, in both its deontic (act-evaluating) form and its telic (world-evaluating) form, by reference to principles that take exactly those modal details into account in determining the permissibility of the acts under scrutiny and completing our pairwise comparisons between worlds.

This Part I tries to accomplish both tasks. First, we situate our nonidentity cases in a modally *enriched* framework rather than a modally *impoverished* framework. And, second, we formulate the intuition itself in modally *sensitive* rather than in modally *constricted* terms.

It may seem that a clear implication of the modal approach is that pairwise comparisons between worlds $w\alpha$ and $w\beta$ may be affected, indeed *reversed*, by events transpiring at still a third world $w\gamma$. Many philosophers will consider that implication highly problematic—and it does sound like I am asking for the recognition of spooky action at a *very* great distance. But we can also put the question in less mysterious terms. It's just whether a modally sensitive formulation

of the person-affecting intuition is ruled out-of-bounds by an *axiological constraint* we have no choice but to accept. A third goal, then, of this Part I is to argue that it's not.

Thus in Chapter 2 below I outline how the most challenging nonidentity cases, as well as the nonidentity problem itself, that is, the argument to inconsistency, are standardly presented. Chapter 3 argues that that standard presentation reflects one or the other or perhaps both of two possible mistakes. Possible Mistake A is the mistake of under-describing, or misunderstanding, the case. Possible Mistake B is the mistake of thinking that how we formulate the person-affecting intuition is ruled out the above-mentioned axiological constraint.

Since the question of whether philosophers have indeed under-described, or misunderstood, their own cases can be settled only by reference to actual cases, I turn in Chapter 4 to what I regard as among the most challenging of the nonidentity cases, that is, Kavka's *pleasure pill case*. We can surely agree that in that case the act is *clearly wrong* and the outcome *clearly worse*. But it's very hard to say *why* the act is wrong and the outcome worse. It might seem that ridding ourselves of the person-affecting intuition would yield a quick solution—that to reject the intuition is to solve the problem. Perhaps that's so (though arguably it isn't). In any case, my argument here will be that a modally enriched understanding of the pleasure pill case, alongside a modally sensitive formulation of the person-affecting intuition, puts us in a position to solve the problem *without* rejecting the intuition. We never get to the result that the act under scrutiny isn't wrong or that the outcome itself isn't worse.

Chapter 5 briefly discusses still another class of nonidentity cases—cases that plausibly avoid Possible Mistakes A and B. I argue, however, that such cases don't meet the *clearly wrong*, that is, *clearly worse*, standard. If a properly formulated person-affecting intuition implies in those cases that the act under scrutiny is *permissible*, it would not be unreasonable to consider the matter closed.

1.2 *Terminology. Possible worlds (futures; outcomes); distributions; accessibility*. We shall suppose that a given history of a given world (such as the

history of the uniquely *actual* world) may unfold in many different possible ways going forward. We will call the many different ways in which such a history of such a world might unfold going forward *possible futures*, *possible worlds* or simply *possible outcomes*.

A *world* is not simply a *distribution*. A *distribution* is a bare-boned description of a world that simply (a) identifies the people that do or will exist at that world and (b) displays the overall lifetime wellbeing levels of each such person. A single distribution may apply to many different possible worlds. But a single world—itself a plethora of detail—determines a unique distribution of wellbeing across a unique population.

We can just note that distributions don't include information as to *other* distributions that agents might bring about in a given case. Worlds, though, are different. The world where agents are able (that is, have the ability) to snap their fingers and eradicate cancer is a distinct world from the otherwise similar world in which agents *lack* that ability. If agents have that ability at a world w1—if, we'll say, a world w2 where agents snap their fingers and cure cancer is *accessible* relative to w1—and agents don't have that ability a world w1′, then w1 and w1′ are distinct.

This last point can be expressed in the form of the *accessibility axiom*.[2] Thus, if in a given case w2 is accessible relative to w1, then in every case w2 is accessible to w1. Agents can't both, in w1, have the ability to snap their fingers and cure cancer and, also in w1, *not* have that ability; worlds are (unlike distributions) far more finely differentiated than that. "Change a fact" about a given world and you have in effect changed the world you are talking about.

Accessible futures are (among other things) possible futures that are not barred by the laws of nature.[3] Not all *possible* futures, relative to a given history

[2] See part 3.6.4.

[3] Whether accessible futures are futures not barred by the acts of other agents is controversial. We might say that futures that aren't accessible to a given agent may nonetheless be accessible to a group of agents and thus consider the future to have a sort of derivative accessibility in respect of the individual agent. What we say on this question has implications for collective action problems, which I won't try to resolve here.

of a given world, are accessible. Thus relative to our own world history—the history, that is, of the uniquely *actual* world—the immediate future in which we snap our fingers and eradicate cancer is *possible*. There's no *logical* or *conceptual* inconsistency in the thought that that particular thwarting of—or sea change in—the laws of nature is right around the corner. But that doesn't mean that the future in which we snap our fingers and eradicate cancer is now *accessible* to us. As things in fact are and, we think, will remain, that particular ability is one we don't have and have no means of acquiring.

Wellbeing; *personal good*; *general good*. *Wellbeing* indicates how good a person's existence at a given outcome is *for that person*. I think of *wellbeing* as that which makes a person's life so precious to the person who lives. For purposes here, we'll have no need to decide whether wellbeing consists in pleasure, happiness, preference satisfaction, resources, capability or something else entirely. But we will need to say a couple of things about wellbeing to proceed. First, a person p's having *more wellbeing* in an outcome wα than p has in an outcome wβ means that wα is *better for p* than wβ is—better for p, that is, "from p's own point of view" (whether or not *recognized* as better for p *by p*). Moreover, a person p's having a *positive* wellbeing level at an outcome just means that p's existence at that outcome is worth having. (Sometimes we'll just say that that person is *happy*.) And p's having a *negative* wellbeing level (p's being *miserable*) just means that p's existence is *less* than worth having—that that existence constitutes a *wrongful life* and that, from p's own point of view, it would have been better never to have existed at all.

The concept of the *personal good*, as distinct from *wellbeing*, will not come into play until Part II. There, we turn to the question whether the existence of an additional happy person in a given outcome makes that outcome better. But I should go ahead and note that *wellbeing* and the *personal good* are two very different things.[4] *Wellbeing* indicates how good a person's existence at an outcome

[4] I use the term *personal good* as Broome does. Broome 2015. Elsewhere Broome calls the personal good *wellbeing*—hence, the meaning he assigns to that term is distinct from the meaning I am assigning to that term here. Broome 2004. As far as I can tell, Broome from 2004 on has no analog to my *wellbeing* even though he makes reference to something's being good *for a person* (mainly

is *for that person*. In contrast, the *personal good* indicates how good a person's existence at an outcome is *for that outcome.* Personal good, in other words, indicates how much value a given person's existence at a given outcome *contributes* to the *overall* value—that is, the *total* good, or what Broome calls the *general* good—of that outcome.[5] We said before that, if p has more *wellbeing* in one outcome than p has in another, then the one outcome is better *for p* than the other is. We can now note that simply in virtue of the meaning of the terms, if p has more *personal good* in one outcome than in another, then (other things equal) the one outcome will be *generally better* than the other.

Recognizing a distinction between *wellbeing* and the *personal good* leaves room for the idea that a person may have a positive *wellbeing* level in a given outcome even though that person's existence in that outcome contributes nothing at all to the *general* good of that outcome. In such a case, *wellbeing* may be positive even though *personal good* is zero. This point will be especially important when we turn, in Part II below, to the question whether the existence of an additional happy person in a given outcome makes that outcome better.

Acts; choices. Often I will use these terms interchangeably; often any act that implements a given choice will have the same morally relevant features as any other act that implements that same choice. In any such case, I will go back and forth freely between the terms *act* and *choice*. However, on occasion we will need to distinguish between *choices* and *acts*. I will then reserve *choice* as the umbrella term and keep in mind that any one of many different possible highly particularized *acts* performed at many different possible worlds may serve to implement a given *choice.* The nature of the agent's *choice* is often decided by the agent prior to the agent's performance of an act that implements that choice. But the nature of the *act* that implements that choice—what the agent ends up *doing*, in all its specificity, at a given world—typically isn't decided (at least, can't be *known*) until

---

to distinguish that from what he does want to talk about, which is the personal good, that is, a thing's being good *for an outcome*).

[5] The terms *personal good* and *general good*, and the relation that is defined between then, come from Broome 2015.

performance is complete. (Compare the *choice* to get a cup of coffee and the particular *act* that will, at the actual world, implement that choice; the *choice* is oblivious to *precise* motion, duration, etc., while the *act's* very identity depends on just those features.)

People. I include as *people* many non-human animals (for example, many mammals, birds and reptiles) in addition to many human beings. But the term also excludes some human beings, for example, human bodies that are alive but whose cerebral cortex is non-functioning.[6] I take for granted that a person who is *never conscious* at a given world *never exists* at that world.[7]

[6] See Peter Singer. [*Animal Liberation*; *Practical Ethics*.] The term *person* thus includes many nonhuman animals and excludes many human beings. For purposes here, I assume consciousness to be a necessary but not a sufficient condition for a thing's being a person.

[7] As just noted, for purposes here, I assume consciousness to be a necessary but not a sufficient condition for a thing's being a person. I assume, moreover, that to survive as the *same* person from one time to another—for the person p at t1 to be numerically identical to the person q at t2—is for consciousness to be knitted together in some fashion or another by a transitive relation of psychological connectedness R. Moreover, I take it that a human or non-human embryo or fetus that hasn't experienced consciousness isn't a person; a human or non-human fetus that *has* experienced consciousness is in close proximity of, but isn't identical to, a person; and the person that may ultimately develop out of a human or non-human embryo or fetus doesn't come into existence until consciousness emerges. Thus: early abortion involves never bringing a person into existence to begin with whereas late abortion might (depending on facts about when consciousness emerges in humans) involve removing a person from existence. Relying on that point, I have argued elsewhere that, while abortion is certainly a matter of killing a fetus, the early or early middle abortion isn't a matter of killing a person but rather of never bringing a person into existence to begin with.

Chapter 2: Standard Presentation and Resolution of the Nonidentity Problem

2.1 *Standard presentation of nonidentity facts*

Population ethics owes much to two basic problems, the *procreative asymmetry* and the *nonidentity problem*. In Part II, I discuss the procreative asymmetry in the context of what John Broome calls the *neutrality intuition*. My focus there will be whether the strongly held intuition that the existence of an additional happy person doesn't, other things equal, make an outcome better leads to inconsistency. The focus of this Part I is the nonidentity problem.

The cases that give rise to the nonidentity problem vary wildly in their specifics. Standard presentations of the facts of those cases, however, track the following outline.

Standard Presentation of Nonidentity Facts (Schematic)

Let w1 be a possible future, or outcome, or *world*—say, the *actual* world, that is, the world as it *actually* unfolds. Let a1 be an act the agent performs at w1. Let p be a child born seriously impaired at w1 as a result of the agent's performance of a1 at w1.

Let w2 be a distinct world, a world available (we'll say *accessible*) to the agent. Let a2 be an act the agent performs at w2 in place of a1 at w1. Let q be a child *nonidentical* to p born healthy at w2 as a result of the agent's performance of a2 at w2 in place of a1 at w1.

We stipulate the following counterfactual: had the agent *not* performed a1 in w1, he or she *would have performed* a2 in w2 and q *would have existed* in place of p.

We stipulate as well that w2 is better for q than w1 is for p (we'll say that p has less *wellbeing* in w1 than q has in w2). We also stipulate that, despite the fact that p's wellbeing at w1 is suppressed as a result of the impairment, p's life at w1 is clearly worth living (p's wellbeing in w1 is clearly in the positive range).

From the fact that p has a life clearly worth living in w1 and never exists in w2, we infer that it's *not* the case that w1 is *worse for* p than w2 is. We also stipulate that no one *other than p* who does or will exist in w1 is affected in any way by what the agent does. Hence we infer that it's *not* the case that w1 is *worse for anyone* who does or will exist in w1 than w2 is.

[END]

Or, in graph form, where bold face means the indicated person does or will exist in the indicated outcome, and italics paired with the asterisk means the indicated person never exists in indicated outcome:

Graph 2.1: Standard Nonidentity Facts (Schematic)		
wellbeing	w1 (including a1)	w2 (including a2)
+10		**q**
+8	**p**	
+0	*q**	*p**

The *problem* arises when we combine the facts of any particular nonidentity case with the *person-affecting intuition*.

2.2 *Standard formulation of the person-affecting intuition; implications*

Let's first focus on the *deontic*, or *act*-evaluating, component of that intuition. The rough idea is that a morally "bad" act performed in a given world must be "bad for," that is, make things *worse for*, at least some person who does or will exist in that world.[8] An act that is bad *only* for the person who never exists at all—in virtue of its leaving *that* person out of existence altogether—cannot be bad, or wrong, at all.

Now, this sensible intuition is often formulated—I would argue reformulated—as the *highly constricted* PAIA(c).[9]

---

[8] See Parfit 1987, p. 363.

[9] Philosophers who have launched the nonidentity problem on the basis of PAIA(c) or an extensionally equivalent principle or set of principles include Shamik Dasgupta (forthcoming), part 1 (combination of claims (2) and (3)). See also Boonin 2014, p. 3 (discussion of premise "P2") and p. 52ff. (Chap. 3). ("The Counterfactual Account is the commonsense account of harm." (Boonin, p. 52)) See also Mulgan 2006 p. 8.

I agree with Dasgupta that an act's being "bad for" a given person p—in, we should add, a "morally relevant sense" (Parfit 1987, p. 374)—involves that act's making things *worse for* p. Thus,

PAIA(c): aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and aα performed at wα is "bad for"—that is, makes things *worse for*, p—than things would have been for p had aα not been performed.

I will argue, later on, that PAIA(c) is *unduly* constricted (part 3.4.1 below).

Let's now consider the *telic*, or *outcome*-evaluating, component of the intuition. The intuition here is that a morally "worse" world must itself make things "worse for" at least some person who does or will exist at that world.[10] The telic component, just like the deontic component, is also typically formulated in terms that are *highly constricted*.[11] Thus:

---

in my view, Harman and Shiffrin are mistaken in thinking that an act's being bad for p can be explained in *non-comparative* terms. It doesn't follow, however, that we must adopt Dasgupta's—or Boonin's—*counterfactual* account of when an act is bad for p. After all, as we shall see in what follows, for reasons entirely independent of the nonidentity problem, that account is highly problematic. Rather, we should adopt a *modally sensitive* comparative account, one that determines that an act is bad for a person, not on the basis of what otherwise *would* have happened but for the act, but rather on the basis of what *could* have happened.

[10] See Parfit 1987, p. 370.

[11] See, e.g., Holtug 2010, p. 158. Holtug—and many other philosophers—go even further: w1 is *better* than w2 only if there is a person who does or will exist in w1 and w1 is *better for* that person than w2. See Holtug 2010, p. 158. See also Fleurbaey and Voorhoeve 2015 [in Hirose and Reisner, eds.], p. 102. (I assume that, when the authors write a "social situation cannot be better than another if it is not better for someone," they mean, "better for someone *who does or will exist in that situation*.)

But the principle that one outcome is better than another *only if* there is a person x who does or will exist in the one outcome and the one outcome is better for x than the other outcome is itself subject to instant counterexample. Thus, in cases involving *wrongful life*, where the person's life is clearly less than worth living and we want to say that, for that person, it would have been better never to have existed at all, the outcome that *excludes* that person is the better outcome even though no person who does or will exist in that better outcome such that that outcome is *better for* that person.

See also Arrhenius 2015 [in Hirose and Reisner]. Arrhenius thus explores the principle that an "outcome A is better (worse) than B" only if "A is better (worse) than B for at least one individual in A or B" (p. 111). I take it that principle implies that A is better than B only if A is better for at least one individual *in A*. If so, that means that this principle too is subject to the instant counterexample of wrongful life.

PAIO(c): wα is morally worse than wβ *only if* there is at least some person p such that p does or will exist in wα and wα is *worse for* p than wβ is.

Applied to our nonidentity facts as those facts are standardly presented, those principles tell us both that a1 is *permissible* and that it's *not* the case that w1 is *worse* than w2, that is, that w1 is *at least as good as* w2 is.

I will argue that—like PAIA(c)—PAIO(c) is *unduly* constricted (parts 3.4.3 and 3.5 below). A less constricted version of the principle seems to capture the person-affecting intuition just as well—but is *far* easier to defend.

2.3 *Completing the problem*

So what's the problem? The problem is that we are *quite* confident that many cases that perfectly track the standard presentation involve acts that are clearly *wrong* and worlds that are clearly *worse than* the worlds we are comparing them against.[12]

Thus we agree with Parfit that the choices of depletion and the risky policy are wrong and with Kavka that it's wrong to sell your own future child into slavery or take the teratogenic pleasure pill prior to conceiving a child—even in the case where *refraining* from performing any of those wrong acts means the child whose plight we purport to be concerned about would never have existed—at least, *very probably* would never have existed—at all.[13]

Ditto for our pairwise comparisons of one world against another. We fully accept that the world in which depletion is implemented is morally worse than the world in which conservation is implemented, and that the world where the risky policy is implemented is morally worse than the world where the safe policy is implemented. And we accept that the world where the parent takes the pleasure

[12] David Boonin disagrees. See Boonin 2014 for the argument that cases that track the standard presentation involve acts that are, after all, perfectly permissible. See also David Heyd 2009 and Heyd 1992 for a metaphysically sophisticated argument reaching the same result.

[13] Parfit 1987, pp. 351-379; Kavka 1982, pp. 93-112.

pill and the one child is born impaired is worse than the world where the parent doesn't take the pleasure pill and a *nonidentical* but better off child is born instead.

Thus we face an inconsistency. We have reasoned our way to the results that a1 *isn't* wrong and that w1 *isn't* worse than w2 is. But we at the time fully accept that a1 *is* wrong and that w1 *is* worse than w2 is.

On the deontic side, the argument to inconsistency can be summed up as follows.

Standard Nonidentity Argument/Deontic Form		
Line no.		*Justification*
1	a1 performed at w1 is morally wrong.	Intuition
2	It's not the case that the act a2 that would have been performed at the world w2 had a1 not been performed at w1 makes things worse for p (or anyone else who does or will exist at w1) than a1 performed at w1 does.	Stipulations (life worth living; counterfactual ; and no one else affected)
3	aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and aα performed at wα is "bad for," that is, makes things *worse for*, p, than things would have been had aα not been performed.	PAIA(c)
4	It's not the case that a1 performed at w1 is wrong.	Lines 2 and 3
5	a1 at w1 both is and isn't wrong	Lines 1 and 4

To solve the problem is, in part, to avoid the inconsistency. But how to avoid the inconsistency?

The nonidentity argument in its telic form proceeds in parallel and similarly ends in inconsistency. Our moral instincts tell us that w1 *surely is* worse than w2—that's line (1). But PAIO(c) applied to the standard presentation of the facts instructs that w1 *isn't* worse than w2—that's line (4).

2.4 *Relation between deontic and telic nonidentity arguments*

I consider the nonidentity problem a problem for *both* the deontic *and* the telic components of the person-affecting intuition.

That way of thinking about the nonidentity problem doesn't seem to be a foregone conclusion. Thus many philosophers focus exclusively on the deontic argument. Those philosophers either (i) may consider the nonidentity problem to directly challenge *just* the deontic component of the person-affecting intuition or (ii) may not consider the project of ranking worlds in terms of their overall betterness a part of moral philosophy (they may, that is, be non-consequentialists). Philosophers subscribing to (i) may consider the telic intuition vulnerable to still *other* challenges but not clearly to the nonidentity challenge. But those same philosophers may instead leave the question of the telic intuition open even as they conclude that the nonidentity problem disproves the deontic intuition outright.

I believe, however, that the two discussions can't effectively be separated. Thus, as we shall see, the solution to the deontic form of the nonidentity problem that I will propose in what follows *requires* that we also, in effect, solve the telic form of the nonidentity problem as well.[14]

2.5 *Standard solution to the nonidentity problem*

The standard solution to the nonidentity problem has been to trace the inconsistency back to a strongly held intuition—that is, the person-affecting intuition—and then to reject that intuition.

The standard solution in the case of the procreative asymmetry has been to do the same—to trace the inconsistency back to the happy child half of the asymmetry; that is, the intuition that (other things equal) the existence of an additional happy child doesn't make the world better—and then to reject that intuition. That intuition itself is an implication of the *neutrality intuition*, which states that the existence of an additional happy person is morally *neutral*. Broome, correctly, argues that the neutrality intuition (understood as the neutral *range* claim)

[14] This argument is made in Chapter 5 below.

is false. He then concludes that that result (somehow) means that we should question—or outright reject—the happy child half of the asymmetry as well.

In Part II, however, I will argue that we can retain the happy child half of the asymmetry while avoiding Broome's objection by replacing the neutrality intuition with what I called the intuition of *narrow neutrality*. I will argue that narrow neutrality—the intuition *behind* the intuition; the intuition that the existence of the additional happy person does not make the world *better*, leaving open that such an addition may easily make the world *worse*[15]—is the only version of the intuition we have any interest in defending to begin with and that it's indeed a version of the intuition we can successfully defend. Through an inversion of Broome's concept of the *personal good*, we can even show that the intuition itself complies nicely with a certain constraint that Broome himself finds compelling, that is, Harsanyi's theorem.

The defense of the person-affecting intuition I propose in this Part I proceeds along similar lines. We can agree that highly constricted formulations of the intuition—such as PAIA(c) and PAIO(c)—fail. But those principles never accurately reflected the intuition to begin with. The problem with them is that they allow us to evaluate the case without ever attending to modal details inherent in those cases—details that may themselves have been, whether recognized or not, at the very root of our unwavering confidence that the act under scrutiny is clearly wrong and the one world clearly worse than the other. In contrast, by attending to the critical modal details of our own cases, and by formulating the intuition itself in modally sensitive rather than in modally constricted terms, we may give ourselves some chance of solving the nonidentity problem *without* abandoning some of what we were pretty sure we knew, that is, the person-affecting intuition itself.

[15] Thus narrow neutrality gives us the room we need to endorse the *second* half of the asymmetry—the intuition that the existence of the *miserable* child (other things equal) *does* make an outcome *worse*.

2.6 *Shared axiology*

Not coincidentally, the intuitions philosophers are quick to put on the chopping block—the person-affecting intuition and the happy child half of the asymmetry—are closely related. Both describe a necessary condition on when an act is wrong or an outcome worse: things must be made *worse for* some *existing or future* person.[16] The necessary condition is left unsatisfied in the case where the happy person in w2 happens to be an *additional* person (in what Parfit calls a "different number" case) and in the case where the *really* happy person q who takes the place of the *merely* happy person p happens to be *distinct* from q (in a "different people" case). With no one else there to satisfy the necessary condition, that condition is then failed. And we then obtain the problem results that the act we are confident is wrong is permissible and the world we are confident is worse is at least as good as the other.

The intuitions seem to have in common an underlying axiology. They both take the value of having more happiness than a world otherwise would have had but having that happiness stuffed into the container of an *additional* person or a *distinct* person—either way, a person q *nonidentical* to a person p—and discount that value to zero.

When, as a result of the nonidentity problem, philosophers reject that particular sensibility, they may consider the way cleared for the *impersonal* view that that value *cannot* properly be discounted: that *more happiness matters morally, regardless* of whether it's stuffed into the container of a nonidentical person or not.

[16] The person-affecting intuition and the happy child half of the asymmetry have much in common. In my view, however, *assuming that the intuitions are suitably formulated*, what they *don't* imply or have in common is the thought that, somehow, it's the people who do or will exist who matter morally, and not the people who will never exist at all. The view that some people matter morally and others not at all, whether articulated in the form of *strong* or *weak moral actualism* (see Hare 2007), does not stand up to scrutiny. The better view is that some diminutions in wellbeing—some *losses*, we might say—matter morally, while others matter not at all. For discussion, see Roberts 2010, Roberts 2011(a) and Roberts 2011(b).

That impersonal approach then is often translated into an *aggregative* approach—into the view that more happiness *in the aggregate* matters morally.[17] Some philosophers—including Temkin—fully accept that *other* values—e.g., equality—matter as well. But the lynchpin to solving the nonidentity problem, in Temkin's view, is to include the maximization of happiness *in the aggregate* as a value. That's the value we just can't get away from (in their view).

But of course it's *widely* understood that theories that abandon the person-affecting intuition—theories that insist on wrongdoing when all the agent has done is *not* bring an additional happy person into existence; theories that insist on wrongdoing when all the agent has done is *not* increase wellbeing on an *aggregate* basis—come with their own deficiencies.

Thus consider *totalism*, the paradigm example of the impersonal, aggregative approach. Many philosophers *like* the account totalism provides of the nonidentity problem.[18] (Totalists ask, "What problem?") But totalism is *widely*

---

[17] Nils Holtug is an exception. Holtug 2010. Holtug shows that we do not need aggregation to solve the nonidentity problem. See Holtug 2010 chaps. 6 and 9.

Holtug describes his approach person-affecting in nature. However, his view—as he points out—takes a wide, rather than a narrow, person-affecting form. He thus abandons the importance of *identity* to moral evaluation. According to Holtug, one outcome's being worse than another isn't necessarily a matter of that one outcome's being *worse for* any *particular* person. Rather, it's a matter of *either* the one outcome's being worse for one person *or* the other outcome's being better for *a possible distinct* person. Holtug 2010, p. 160.

When it abandons the identity condition, Holtug's view abandons the person-affecting intuition in what I take to be its most interesting form. It's that form of the intuition that I want to defend here against the nonidentity problem.

Fleurbaey and Voorhoeve (2015 [in Hirose and Reisner, eds.], pp. 103-105) similarly avoid aggregation but also abandon the identity condition.

[18] Not all philosophers think totalism does a good job with the nonidentity problem. Thus Elizabeth Harman argues that any satisfactory solution to the nonidentity problem will not simply generate the result that the act under scrutiny is wrong but also will provide an explanation of the wrongness of that act that is rooted in *what has been done to the child* we intuitively consider the victim of that wrong act. See Harman 2004, p. 90.

Holtug 2010 and Dasgupta (forthcoming) seem to agree with Harman that an adequate explanation needs somehow to reference the child whose plight has triggered our concern. At the same time, though, on their views it's not what has been done to *that child alone* that accounts for

understood to leave us fumbling badly when we consider what it has to say about some of the *many other* problems that arise in population ethics. Those include the *repugnant conclusion*, the *replaceability* problem and the *infinite population* problem. [19]

Sophisticated extensions of totalism do not do much better. Critical level theories must worry about the *sadistic conclusion*. And hybrid, or pluralistic, theories have a hard time generating any clear results at all and thus a hard time gaining our support. How can we accept a theory we *can't test*?[20]

---

the wrongdoing. Rather, it's also, in part, the fact that the one child is worse off than an alternate, better off child would have been.

[19] To see how difficulties with totalism—and averagism as well—may have motivated interest in developing an alternate approach, one that focuses on the "happiness of individuals rather than happiness on the whole," see Lazari-Radek and Singer 2014, pp. 363-73.

See also Parfit 1987, pp. 381-390. Most philosophers are completely nonplussed by the accounts totalism provides of the repugnant conclusion, replaceability and the infinite population problem. But some philosophers, I should note, bite the bullet and simply accept even what seem to be totalism's most unfortunate results. Thus, for an argument that we should accept the repugnant conclusion, see e.g. Huemer 2008.

[20] Larry Temkin takes the impersonal approach as far as he can in solving the problems of population ethics. Thus he describes what he calls the *internal aspects* view which recognizes the maximization of wellbeing in the aggregate (let's call this *additive maximization*) as a value but recognizes myriad other (impersonal) values as well (*equality*; *human flourishing*). Additive maximization means that the internal aspects view can offer a straightforward solution to the person-affecting person. But in the end Temkin makes a compelling case that the *essentially comparative* view—which itself recognizes a certain formulation of the person-affecting intuition—cannot be entirely set aside. *Both* approaches are, Temkin argues, in the end going to be called on if we are ever to solve the full range of problems in population ethics.

Hence what I call Temkin's *radical pluralism*. The testing challenges Temkin's theory faces are, however, profound. Not only does application of theory require that we balance many different *impersonal* values against each other (additive maximization against equality against human flourishing, etc.). It also requires that we balance all those *impersonal* values against *person-affecting* values. Temkin himself would be the first to acknowledge that the contemplated balancing procedure is itself going to be difficult to define and hence to test. See *Rethinking the Good* (Oxford 2012).

Depending on how we understand Temkin's overall argument, another question arises. If a theory X implies a world wα is at least as good as a world wβ is, and everything in X is true, then as a matter of logic X in combination with a theory Y also implies that wα is at least as good as wβ is. If knowing *less* supports a give result in a valid scheme, then knowing *more* supports exactly that same result. That means that the argument that (1) just focusing just on wα and wβ leads us to conclude, given X, that wα is at least as

Both intuitions—both the person-affecting intuition and narrow neutrality—may thus have a critical role to play in moral evaluation. It thus makes sense to look hard at the arguments that purport to show that those intuitions fail before we cast them aside.

---

good as wβ is, but (2) then taking wγ in account leads us to conclude, now given X *and Y,* that wα *isn't* at least as good as wβ is (that wα is worse than wβ is), can't work. We need to reject (1) and say instead that we didn't validly conclude, given X, that wα is at least as good as wβ is; we need to say instead that we need to look beyond wα and wβ to complete the comparison.

If that is indeed Temkin's overall argument, then the right conclusion would not be that the internal aspects view and the essentially comparative view are both true and represent values that need to be weighed against each other, but rather than the internal aspects view is false.

Chapter 3

Critique of Nonidentity Argument

3.1 *Two possible mistakes*

I believe that the nonidentity argument to the result that a1 is both permissible and wrong fails. Ditto the argument to the result w1 both is and isn't worse than w2. And I believe that the widespread perception that the arguments succeed is rooted in one or the other or perhaps both of two serious mistakes.

The first possible mistake has to do with how philosophers have presented their own cases.

> *Possible Mistake A.* Philosophers have failed, starting out, to present the facts of their nonidentity cases—Depletion; Risky Policy; Slave Child; Pleasure Pill and many others—at an appropriate level of detail. In situating their cases within a *modally impoverished* framework—a framework that obscures or leaves unmentioned critical *modal details* inherent in their cases—rather than a *modally enriched* framework, they may have, in effect, under-described or perhaps even misunderstood the nature of their own cases.

Possible Mistake A creates two sorts of risks. First, the modal details that are left out of the under-described case might be details that are *critical* to solving the problem—*critical*, in particular, to avoiding the inconsistency *without* abandoning the person-affecting intuition.

Second, when critical details are left unrecognized, philosophers may be misled into thinking that still *other* details of the case *must* be *terribly telling* when in fact they are red herrings and tell us nothing at all. We basically start making things up when we give ourselves too little to work with starting out. We then design solutions that make *non*-critical details their centerpiece and thereby confine our solutions to ones that *can never actually work.*

The second possibility is this.

> *Possible Mistake B.* Philosophers have formulated the person-affecting intuition in terms of *highly constricted* principles—principles like PAIA(c) and PAIO(c); principles that, as we shall see, make the above-mentioned critical modal details irrelevant to the moral evaluation. I take for granted that philosophers have had some reason for going down that path. Perhaps, for example, they

> think that the person-affecting theorist is bound by an *axiological constraint* that immediately rules out-of-bounds *modally sensitive* formulations of the intuition. But that assumption would itself be a mistake if, e.g., that axiological constraint doesn't have the reach or strength it's assumed to have.

If *that's* what's going on, then it really doesn't matter whether philosophers carelessly situate their cases within *modally impoverished* frameworks or conscientiously within *modally enriched* frameworks. Either way, they will consider the *axiological constraint* to bar any account that would bring those modal details to bear in solving the problem.

According, then, to Possible Mistake A, philosophers mistakenly think that *they* haven't arbitrarily limited the range of principles available to them for purposes of solving the nonidentity problem but rather that that range *just is limited* by the facts of their own cases. According to Possible Mistake B, philosophers mistakenly think that there's no point carefully ferreting out the modal details of their own cases since an axiological constraint will in any event make those details irrelevant for purposes of the moral evaluation.

I won't try to figure out here which possible mistake—A or B or both; missing facts or bad principles or both—is more probable. Rather, my purpose here is to show that avoiding both mistakes can help us make progress in solving the nonidentity problem *without* abandoning the person-affecting intuition.

3.2 *Possible Mistake A: Missing facts*

We might make the mistake of under-describing our own case. The standard presentation of the nonidentity case—Graph 2.1—makes just that mistake when it fails to specify whether the accessible outcomes for the agents are exhausted by w1 and w2. Consistent with that presentation, there may well exist some third accessible outcome w3 such that p is better off in w3 than p is in w1. If there is such a better-for-p w3, then it will also be true that, while w1 *isn't* worse for p than *w2* is, w1 *is* worse for p than *w3* is—and thus worse for p than *some other accessible outcome* is.

Now, not *all* cases that are thought to give rise to the nonidentity problem take that form. I believe, however, that the cases that give rise to the *most challenging* form of the nonidentity problem—cases, that is, in which the act under evaluation is *clearly wrong* but it's very hard to say *why*—take *exactly* that form. They are, that is, not *two outcome* cases, but rather *three outcome* cases.

That claim, of course, requires an argument. But it's not an argument that has any persuasive force made in a vacuum. We'll need to examine an actual nonidentity case. And that we shall do when we turn to Kavka's pleasure pill case (Chapter 4 below).

But in the meantime it's important to note that to accept that claim—to accept that the most challenging forms of the nonidentity problem are based on *three outcome* cases—is *not* to solve the nonidentity problem. Indeed, a number of philosophers accept, or seem at least prepared to accept, that their own cases include a better-for-p third accessible outcome w3.[21] They just think that the bare fact that w3 is *accessible*—that it's not at odds with the laws of nature; that it's something agent have some remote chance of bringing about; that it's *technically* accessible—doesn't take us very far at all in solving the nonidentity problem. I agree. The nonidentity problem is far more interesting than that.

3.3 *Setting aside third accessible outcome as irrelevant*

Thus philosophers who seem to accept a better-for-p third accessible outcome often nonetheless set that third outcome aside as irrelevant to the moral evaluation. Some philosophers offer one basis for that set-aside. Others at least suggest a second.

3.3.1 *Probability set-aside*. Philosophers sometimes offer *probability* as reason for analyzing a given nonidentity case as though it were a two outcome rather than a three outcome case. They may concede that some such better-for-p

[21] I include Kavka and Parfit here. Thus Kavka explicitly recognizes that agents might bring it about that one and the same child is better off, and Parfit asks us to accept that, after a few generations, depletion would in fact yield an entirely non-overlapping population but never asks us to accept that any such overlap is impossible.

w3 exists as an outcome that is *technically* accessible to agents. But they then push back: conceding w3 is accessible, they then argue that it's highly *improbable—so* improbable that for the purposes of analysis it can be ignored. Surely, even if agents *tried* to bring w3 or anything like w3 about—had agents, that is, *tried* to make things better for p than things are in w1—their chances of *success* would have been very close to zero. Very probably, they instead would have ended up leaving p out of existence altogether.

Kavka thus described the *precariousness* of the coming into existence of any particular person. Had agents done things otherwise than *just as they did* in the period of time leading up to the conception of the *particular* person p, they would thereby have affected—changed—at least in some slight way the *timing and manner* of conception.[22] And affecting the timing and manner of conception in even the slightest way would surely have reduced *p's* chances of coming into existence to almost nothing at all. (It's the males, mainly, who are behind the precariousness of existence, producing 200 million plus sperm cells per sexual encounter, a distinct inseminating sperm cell, it seems reasonable to suppose, yielding a distinct person.)

From that probability point, philosophers segue to the position that for all practical purposes, including moral evaluation, we might as well just ignore w3—that we may as well analyze the particular nonidentity case as a two outcome rather than a three outcome case.

But that segue isn't valid. We are familiar with and should accept the logic that (presumably) is *supposed* to support the inference. Consider the following medical example. If Ginger is sick and facing her own imminent demise, she prefers a treatment X that promises a very high probability of survival *and* a very good life if she *does* survive in place of a treatment Y that promises that, if she survives, her life would be *even better* than "very good" but also promises almost *no chance of survival at all*. Ginger, correctly, considers treatment X so clearly better for her that she doesn't think there's any *practical* reason at all for even

[22] See Kavka 1982, p. 93; Parfit 1987, p. 361.

keeping treatment Y on the table as an alternative worthy of any further serious consideration.[23]

We can and should accept that logic. The problem is that it doesn't apply to the most challenging versions of the nonidentity problem.

That the low probability of the agents bringing about the better-for-p *w3* means that we can set w3 aside as irrelevant for purposes of our evaluation is a mistake *unless* we *also* happen to know that the probability of the agents bringing about the not-quite-so-good-for-p *w1* is somehow *greater*.

Thus it's *stipulated* in the medical example that treatment X would give Ginger a *very high probability* of survival. But nothing in the most challenging versions of the nonidentity problem tells us that the probability of the agents bringing about w1—or any outcome in which p exists and has a very good life—is itself *very high*.

It might be argued that w1's high probability (versus w3's low probability) is just *presumed* as a fact of the case—a fact we are absolutely free to make perfectly explicit in the standard presentation if we happen to suffer from OCD and thus feel ourselves compelled to spell out such an obvious point.

But how can we *presume* that the precariousness of existence applies to p's coming into existence given *a3* but (somehow) *doesn't* apply to p's coming into existence given *a1*? Why should we think that the *wrong act* a1 will somehow make it more probable—measured at the appropriate moment; that is, the moment just prior to performance—that p will eventually come into existence than will at least *some* other act that we take to be *better* for p, e.g., a3?

If it's then correct in the particular case that the precariousness of existence cuts both ways—that the probability of p's coming into existence is very low whether a1 is performed or a3 is presumed; that probability of w1, given a1, or, more accurately, given a choice that *includes* a1, is just as low as the probability of w3, given a3, or, again, a choice that *includes* a3[24]—then the fact that the better-for-p w3 exists as an accessible outcome becomes *highly significant*.

[23] [add cites]

To see that that's so, we just need to slightly revise the medical example: if you are ill and the probability of cure, miniscule under Y, is equally miniscule under treatment X, then of course you prefer treatment Y. We all prefer (other things equal) the very small chance of winning the *billion* dollar lottery over the same very small chance of winning the *million* dollar lottery.

This point can be put in terms of *expected wellbeing*, where the *expected wellbeing* of a given act for a given person is just the summation, for each possible outcome of that act, of the actual wellbeing assigned to that person by that outcome multiplied by the probability that that outcome will obtain, given that act. Thus the *expected wellbeing* X produces for you in the second version of the medical example is greater than the *expected wellbeing* that Y produces for you. Ditto the lottery example. The *expected wellbeing* produced for me when I pay a dollar for a very small chance at a billion dollar jackpot is greater than the *expected wellbeing* produced for me when I pay the same dollar for the same very small chance at a million dollar jackpot.

This is not, of course, to argue that *all* future-directed acts share the identical, *very low* probability of bringing any particular person into existence. We can easily construct a case in which the probability of w1 *is* greater than the probability of w3—or, more precisely, where the probability of the agent achieving the lesser outcome for p is greater than the probability of the agent achieving the better outcome for p. And all we need, for purposes of showing that the person-affecting intuition is false, is *one* successful counterexample. Effective, but risky, fertility treatments—treatments that increase the chances of conception but also increase the chances of the child's being burdened in some way if he or she is conceived—come to mind. Depending on the actual numbers, such treatments may well generate more expected wellbeing for a particular child than a lower-risk but largely ineffective alternative. But is it really so clear—provided the risk itself is

[24] More accurately: if the probability of p's coming into existence is very low whether the agent makes the choice that includes a1 or the choice that includes a3—if, that is, the probability of w1, given the choice that *includes* a1, is just as low as the probability of w3, given the choice that *includes* a3—then w3's accessibility becomes highly relevant to the analysis, notwithstanding its low probability of obtaining.

minimized (that is, that there is no still better alternative for the child) and that there's no chance of anyone else being affected in any way—that it's wrong for the physician to provide or for the patient consent to the effective, high-risk treatment? I don't think so.

The upshot here is that we can't presume the probabilities one way or the other in advance of our close scrutiny of the details of the particular case. We certainly can't *presume* that the phenomenon of the precariousness of existence *only* applies when the act under scrutiny is one we consider *clearly permissible* and that it somehow disappears into the woodwork when the act is one we are confident is *clearly wrong*. Nor can we presume that the probabilities are always a wash. We shall instead have to look at the cases.

I think then what we will find—and that is the topic of Chapter 4—that the cases that pose the only *clear* challenges to the person-affecting intuition—the cases, that is, in which the choice *is* clearly wrong; and here I include depletion, risky policy, slave child, pleasure pill, historic injustice and many others—are cases in which the (very low) probability of the agent's achieving the not-quite-so-good-for-p w1 on behalf of p is *no greater than* the (very low) probability of the agent's achieving the better-for-p w3 on behalf of p.

3.3.2 *Counterfactual set-aside*. Other philosophers may set w3 aside as irrelevant for reasons having nothing to do with probability. They may think that w3 is irrelevant because they consider the following counterfactual highly relevant to the analysis: had agents refrained from performing a1 and thus failed to bring about w1, then the agents *would have performed* a2 at w2 instead. They may think, in other words, that in view of that counterfactual what's going on at *w2*—but *not* what's going on at *w3*—is relevant to the discussion.

Exactly that counterfactual is included as a stipulation in the standard presentation of the nonidentity case.

But it's a mistake to think that that counterfactual makes w3 irrelevant to the evaluation. That I *would* have done still worse to a given person p, had I not just as I did, is never a vindicator of what I have in fact done *unless* it's also true

that I had no third, better-for-p alternative. When I do have such a third, better-for-p alternative, whatever I *would* have done, what I *have* done may well still be wrong.

Suppose, for example, that I shoot Harry in the arm and that, had I not shot him in the arm, I would have shot him in the heart. Surely, in that case, my shooting Harry in the arm is permissible *only if* I somehow don't have the third alternative of just standing there and not shooting Harry at all. It's the existence or non-existence of that third alternative that determines whether my shooting Harry in the arm is permissible—*not* the fact that I *would* not have availed myself of that alternative had it existed. (The *world* decides what is permissible, not what happens to be convenient for the *agent*.)

3.4 *Possible Mistake B: Bad principles*

Here, we must worry about whether the *deontic* and *telic* components of the person-affecting intuition are properly formulated.

3.4.1 *Bad deontic principle.* According to the deontic, or act-evaluating, component of the person-affecting intuition PAIA(c), aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and aα performed at wα makes things *worse for p* than things *otherwise* would have been—than things, that is, would have been *but for* wα.

> PAIA(c): aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and aα performed at wα is "bad for"—that is, makes things *worse for*—p than things would have been had aα not been performed.[25]

By directing us just to compare what agents *have done* to a given person against what those agents *would have done* to that person had agents not done the thing they *have* done, the highly constricted PAIA(c) obviates any need to guard against Possible Mistake A. By its very nature, PAIA(c) makes any accessible outcome beyond wα and wβ irrelevant to the analysis.[26]

---

[25] For philosophers who have their discussions of the nonidentity problem with a formulation of the person-affecting intuition very like PAIA(c), see note 8 above. Parfit is an exception.

[26] There are, of course, other unfortunate formulations of the person-affecting intuition in addition to PAIA(c). One such formulation noted and then rejected by Lazari-Radek and Singer, and previously Singer, is the *prior existence view*. Lazari-Radek and Singer 2014, p. 368; Singer 2011, pp. 88-89. They argue, convincingly, that the miserable child half of the procreative asymmetry shows that the prior existence view will not work.

More generally, the miserable child case and many others convince us that *any* principle that tries to draw a distinction between people who have moral status and people who don't—on the prior existence view, the miserable child doesn't have moral status in virtue of the fact that that child isn't among those people who do or will exist *however* the choice under scrutiny is made—will fail. That would include views that deem only *actual* people (whether existing or future) to have moral status and views that deem only the people who do or will *exist under the act under scrutiny* to have moral status. For more on *moral actualism*, see note [15] above.

But not all formulations of the person-affecting intuition make that mistake. PAIA* and PAIA** don't. They avoid the result that it's permissible to bring the miserable child into existence, and they at the same time imply that it's not wrong not to bring the happy child into existence.

Philosophers who think that the person-affecting intuition is to be understood by reference to PAIA(c) then have little difficulty convincing us that PAIA(c) generates false results in many nonidentity cases.

The difficult with their argument is that there is no good reason to think that PAIA(c) captures the person-affecting intuition—that the necessary condition expressed by the person-affecting intuition is as stringent; that is, as easily failed—as PAIA(c) says that it is.

Thus, the person-affecting intuition is best understood *not* to imply that what the agent has done is permissible *whenever* what the agent *would have* done, had the agent not done just what the agent did, makes things *still worse* for the person. Rather, it's best understood to imply permissibility just when the agent had *no way at all* of making things *better* for that person.[27] After all, we can surely agree that PAIA(c) is simply false.

We needn't mine the area of population ethics in order to see why that is so. The shoot-Harry-in-the-arm case does that work for us. It's the fact that I have a third alternative in that case—the alternative of just standing there and not shooting Harry at all—that shows that my shooting Harry in the arm was wrong.

And there's a larger lesson here as well: we can't accurately evaluate my shooting Harry in the arm until we have *looked around at all the details of our own case*—including the *modal* details; the details regarding what *could* have been—and noted that I had the alternative of not shooting at all. Had I not had that alternative, my shooting Harry in the arm would have been permissible, indeed, obligatory.

If that point holds for the easy, ordinary, *same* person case, then why it would not also hold for the harder, extraordinary, *additional* person case is very unclear.

Critically, however, those same principles *are* vulnerable to the *nonidentity problem*. Hence the motivation for this Part I.

[27] The person-affecting intuition is best understood, in other words, to open the door to a finding of wrongdoing whenever agents fail, for each person who does or will exist, to *maximize* wellbeing for that person.

It seems, then, that we should formulate the deontic component of the person-affecting intuition by reference not to the *highly constricted* PAIA(c) but rather to the *modally sensitive* PAIA*.

> PAIA*: aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and there is an alternative act aβ performed at an alternative accessible outcome wβ such that aα performed at wα makes things *worse for* p than aβ performed at wβ does.[28]

In other words: if aα performed at wα *maximizes* wellbeing for each person who does or will exist at wα, then aα *isn't* wrong.

PAIA* in hand, I can now outline the response to the nonidentity problem that I think works. For the most challenging forms of the problem—those in which the act under scrutiny is *clearly wrong*—we regularly find that we are perfectly able to identify a third, better-for-p alternative, an alternative beyond what the agent in fact does and beyond the counterfactual alternative the critic of the person-affecting intuition wants to restrict our attention to. We then simply note that, in such a case, PAIA* avoids the implication of permissibility and thus opens the door for a finding of wrongdoing based on still other, widely accepted, person-affecting principles.[29]

3.4.2 *Objection based on probabilities*. It might seem that PAIA* fails to capture an important element of the person-affecting intuition. That PAIA* itself sidesteps a finding of permissibility in a given nonidentity case is not very useful if

[28] I am happy to say that, if there is such an act aβ at wβ, that aα *harms*, or imposes a *loss* on, the person it makes things worse for. It would be a mistake, however, to get bogged down in the meaning of the term *harm* or *loss* and hence my use of those terms may be regarded as simply shorthand for making things *worse for* a given person in the sense described by PAIA*.

[29] Pareto principles are good examples here. Thus where w3 is accessible relative to w1, and w1 and w3 contain the same people, and w3 is better for at least one person and worse for none, we will say that the act the produces w1 is itself wrong. For discussion, see Roberts 2010 (*Abortion and the Moral Significance of Merely Possible Persons*).

some other important element of the person-affecting intuition generates exactly that result.[30]

Specifically, it might seem that PAIA* is blind to the above-mentioned phenomenon of the *precariousness* of existence. It's one thing to require, for wrongdoing, that a better-for-p alternative exist as a *technically accessible* outcome. That requirement is—we might well agree, at least for the most challenging versions of the nonidentity problem[31]—relatively easy to satisfy. And indeed my argument—in Chapter 4 below—will be that for those versions of the problem the requirement *is* satisfied.

But surely the person-affecting intuition isn't just about *technical accessibility*. Surely it also has something to say in the case where the agent can *technically* bring about the better-for-p outcome but where, *whatever* the agent does, the *probability* that that better-for-p outcome will obtain is *very, very low*. In other words, surely the person-affecting intuition includes the idea that an act—say, a1 at w1—is wrong *only if* agents have *at least as good a chance* of bringing about the better-for-p w3 as they have of bringing about the not-quite-so-good-for-p w1. Roughly put, the idea here is this: only when the alternate, technically accessible, better outcome is an outcome agents had *some significant chance* of bringing about do we have room to declare that the one act leading to the one outcome is *wrong*.

Now, the *actual* value person-affecting theorist might resist that point, taking the position that PAIA* exhausts the deontic component of the person-affecting intuition. But many philosophers want their moral principles to be *action-guiding*. The question is whether a person-affecting complement to PAIA* can be

---

[30] [where to place?] We noted earlier that the fact that a better-for-p accessible outcome is *highly improbable* doesn't mean that we should deem that outcome *irrelevant* for the purposes of moral evaluation. Rather, the legitimacy of setting aside any such improbable better-for-p outcome as irrelevant will depend on whether any not-*quite*-so-good-for-p alternative accessible outcome is any *less* improbable. I think we put that particular point to bed at least for purposes of *theory*. As a *theoretical* matter, that point is indisputable.

[31] There are plenty of other nonidentity cases in which it's failed. In my view, however, those cases are not among the most challenging. For further discussion of this other sort of nonidentity case, see Chapter 5 below.

formulated that can play that that action-guiding role *without* falling prey to the nonidentity problem.

What we need, then, is a principle—we'll call it *PAIA*** in what follows—that considers *expected* wellbeing—not just *actual* wellbeing—to have a role to play in determining permissibility.

But in designing PAIA** we should insist on one condition—the *level playing field* condition. We should insist that analysis under PAIA** *not* be weighted in *favor* of the act under scrutiny—that is, the wrong act—and *against* each alternative to that act. Thus, if we decide that probabilities are important, and if, on that basis, we decide that what is relevant to our moral evaluation of a1 is the *expected* wellbeing generated for p by *each alternative to a1*, then so must we be committed to the position that what is relevant to our moral evaluation of a1 is the *expected* wellbeing generated by *a1*.

Specifically, we should resist the thought—tempting as it may be, in a *post hoc ergo propter hoc* sort of way, given that how the future *actually* unfolds at w1 is *stipulated* as part of the case; so tempting that what we really face here might be called the nonidentity *fallacy*—that evaluating a1 is just a matter of comparing the *actual* wellbeing a1 generates for p at w1—very high, since p has a very good life in w1—against the *expected* wellbeing each *alternative* to a1 generates for p at each alternative accessible world—*very* low, given the very long odds against any one person ever being conceived at all *however* agents comport themselves. Rather, it's the *expected* wellbeing of a1 for p that should be compared against the *expected* wellbeing of each alternative to a1 for p.[32]

Thus, according to this new principle, if aα performed at a world wα *maximizes expected wellbeing* for each person who does or will existing in wα, then aα *isn't* wrong.

---

[32] For those philosophers who *don't* find it reasonable to think that probability has a role to play in determining permissible, we would be happy with the actual wellbeing against actual wellbeing comparison. The point here is that what won't do is mixing apples and oranges in this context.

I have thus elsewhere argued that an approach that determines moral permissibility on the basis of a comparison between one act's *actual* value and another act's *expected* value is inconsistent. The point is an obvious one but widely disregarded in the nonidentity literature.

PAIA** expresses that idea.

> PAIA**: aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and there is some alternative act aβ performed at an alternative accessible world wβ such that the expected wellbeing of aα performed at wα for p is less than the expected wellbeing of aβ performed at wβ for p.[33]

Accepting PAIA** as an element of the person-affecting intuition raises the stakes. It means that for the most challenging forms of the nonidentity problem—those in which the act under scrutiny is *clearly* wrong—to defend the intuition I will need to show that we really can identify an alternative act performed at an alternative accessible world that satisfies the necessary condition on wrongdoing that PAIA** provides. I will need to show, that is, that we can identify an alternative act that generates still more *expected* wellbeing for the person whose plight has concerned us starting out than the act we agree is clearly wrong does. I turn to that work in Chapter 4 below. But first we need to consider the telic, or outcome-evaluating, component of the person-affecting intuition.

3.4.3 *Possible Mistake B: Bad telic principle*. Philosophers often formulate the telic, or outcome-evaluating, component of the person-affecting intuition in the form of PAIO(c).

> PAIO(c): wα is morally worse than wβ *only if* there is at least some person p such that p does or will exist in wα and w1α is *worse for* p than wβ is.[34]

But why should we think that the telic component of the intuition must be forced in the mold of the highly constricted PAIO(c)?

---

[33] For my own part, I would accept both PAIA* and PAIA**—accept, that is, that, for an act at a world to be wrong, there must be some person p who does or will exist at that world and some alternate act performed at some alternate accessible world such that both the actual and the expected value of the alternate act for p is greater than the actual and the expected value of the one act for p.

[34] For philosophers who have proposed this formulation, see note 10 above.

Why not, that is, immediately abandon PAIO(c) as, e.g., a typo and go straight to PAIO*:

> PAIO*: wα is worse than wβ, only if there is some person p who does or will exist in wα and some accessible outcome wγ (which may but need not be identical to wβ) such that wα is worse for p than wγ is.

In connection with the evaluation of *acts*, the person-affecting intuition is best understood to require us to look around at *all* the facts of our own cases before deciding permissibility rather than focusing exclusively on what *was* done and what otherwise *would have been* done. Why not say the same thing about the evaluation of *outcomes*—that there, too, the person-affecting intuition is best understood to require us to look around at *all* the facts of our own cases before deciding whether any one outcome is morally worse than any other?

To my knowledge, critics of the person-affecting intuition don't explicitly answer (or raise) that question. Rather, they simply formulate the telic component in terms of PAIO(c) or a similarly highly constricted principle and then proceed to the (rather straightforward) task of refuting that principle.[35]

We are thus left to speculate. My own sense is that the answer to the question has little to do with the nonidentity problem or the person-affecting intuition and everything to do with a certain *axiological constraint* that many philosophers, including many critics of the person-affecting intuition, consider applicable to any proper pairwise comparison between outcomes. They, in other words, assume that PAIO* is ruled out-of-bounds by an *axiological constraint* we have no choice to accept. I consider that possibility now.

3.5 *Axiological constraint*

[35] Thus I think we can, and should, reject PAIO(c). But my main goal is to show that we can reject PAIO(c)—which was never an adequate way of capturing the person-affecting intuition to begin with—without rejecting the intuition. We shall simply do a better job articulating the intuition. We should prefer PAIO* to PAIO(c).

3.5.1 *Source of constraint; the classic view*. Since the distinction between PAIO(c) and PAIO* becomes apparent only in the context of three outcome cases—they generate the same result in any *two* outcome case—I focus on the three outcome case here.

Graph 3.5.1: Three Outcome Schematic			
wellbeing	w1 (including a1)	w2 (including a2)	w3 (including a3)
+10			**p**
+5	**p**		
+0		*p**	

It's easy enough to produce an argument that concludes with the axiological constraint. We can do it in two quick steps.

The *first step* is to accept what I will call the *classic view* regarding how any proper pairwise comparison between outcomes is to proceed. The classic view is reflected in totalism. But it's a view that many maximizing consequentialists—including many pluralists—share.

Maximizing consequentialists are of course accustomed to what I would consider *modally sensitive principles* when it comes to the evaluation of *acts*. No one thinks that the fact that aα is better than aβ means that aα is *permissible*.

Things change, however, when we turn to the evaluation of *outcomes*. According to the classic view, the pairwise comparison between accessible outcomes can be completed—not *must* be completed; simply that, in contrast to the case of *acts*, there is no reason the pairwise comparison *can't* be completed since there the goal is not to determine which outcome is *best* but rather to determine which outcome is *better*—by examination of *just* the *two* outcomes and *without* reference to any *third* outcome.

But of course, if the comparison between wα and wβ *can* be completed without examining any third accessible outcome wγ or indeed even knowing whether such a third wγ exists, then that would mean that we don't *need* to look beyond wα and wβ to rank wα against wβ. But if we don't *need* to do that, then

anything beyond wα and wβ *cannot make a difference* to how we compare wα against wβ. And if nothing beyond wα and wβ *can* make a difference to how we compare wα against wβ, then allowing considerations relating to wγ—those *modal details*—to affect—that is, to *change*—how we rank wα against wβ would be irrational.

And it's that last point—that core component of the classic view—that I want to focus on here: allowing considerations relating to wγ—those *modal details*—to affect—that is, to *change*—how we rank wα against wβ would be irrational.

The *second step* is then to squeeze the person-affecting intuition into the classic view—that is, to give the person-affecting intuition the appropriately person-affecting role to play a role in ranking w1 against w2 *but* at the same time making sure that that ranking can be completed *without looking beyond w1 and w2*.

The result is a commitment to the position that that the person-affecting necessary condition on when w1 is worse than w2 can be satisfied *only in a certain way*. Specifically, it's a commitment to the following position:

> *Axiological constraint*. The person-affecting necessary condition on w1's being worse than w2 can be satisfied *only if* w1 is worse for a person p who does or will exist in w1 *than w2 is*.

Thus the *axiological constraint*—a restriction on how the telic component of the person-affecting intuition is to be formulated.

PAIO(c), of course, satisfies that constraint. But it's a constraint that PAIO*—which explicitly takes what is going on at w3 into account in ranking w1 against w2—immediately fails.

Now, we'll need to examine the axiological constraint. Is it really one that the person-affecting theorist has no choice but to accept? We will turn to the arguments in part 3.6 below.

But first we attend to some miscellaneous matters: underlining why the axiological constraint represents such a deep challenge against a person-affecting approach (3.5.2); noting (relatedly) that the options for maintaining a person-affecting approach to the nonidentity problem without discarding the axiological

constraint are unpalatable (3.5.3); and finally articulating the implications the axiological constraint has for the particular version of the neutrality intuition—that is, *narrow neutrality*—I defend in Part II.

3.5.2 *How the axiological constraint undermines a person-affecting solution to the nonidentity problem.* Under plausible assumptions, accepting the constraint means that our person-affecting deontic solution to the nonidentity problem—outlined in Chapter 2 above and filled out in Chapter 4 below—cannot be made to work. Those assumptions include (i) that there exists a *very tight connection* between act evaluation and outcome evaluation; and (ii) that the *at least as good as* relation between outcomes is *transitive*.

Why? The very tight connection between act and outcome evaluation makes the following procedure possible and indeed natural. We first rank the outcomes. We then determine among those outcomes which are accessible. And finally we evaluate the acts that bring those outcomes about. Thus an act that produces a given outcome is wrong *if and only if* that outcome is worse than at least some other accessible outcome is.[36] Applying that procedure to a three outcome case—Graph 3.5.1—we understand that, if w1 < w3 and w3 is itself accessible, then a1 is wrong. Now, while the *procedure* is to rank the outcomes and then evaluate the acts, the *inferences* go in both directions. Thus, if a1 at w1 is wrong, then w1 itself must be worse than at least *some* other accessible outcome.

We then compare w1 against w2. According to PAIO(c), it's not the case that w1 is worse than w2, there being no existing or future person p in w2 such that w2 is worse for p than w1 is. Hence: w1 is at least as good as w2; w1 ≥ w2. But PAIO(c) also instructs both that w2 ≥ w1 and that w2 ≥ w3, there being no existing or future person in w2 for whom w1 or w3 is worse than w2 is.

[36] Some philosophers who question the very tight connection between act and outcome evaluation do so on the basis of the special obligations they think we have in respect of our own nearest and dearest. But I find *impartiality* the more plausible view. See Hirose, "A Puzzle from Nagel's Pairwise Comparison" [in AAA Research]. And I see no reason that we cannot *both* accept impartiality *and* a well-articulated, modally sensitive person-affecting approach—accept, that is, impartiality *but also* accept that identity is important. See note 1 above.

Assuming transitivity, we then infer that w1 ≥ w3.

The problem for the deontic person-affecting solution to the nonidentity problem may now be obvious. For that solution to work, a1 at w1 can't be—and wasn't, under PAIA*—deemed permissible. Avoiding the implication of permissibility opens the door to still other person-affecting principles that would evaluate a1 at w1 as *wrong*.[37] But if a1 at w1 is wrong, then the *tight connection* between act and outcome evaluation means that w1 must be *worse* than at least *some* accessible outcome. Here the obvious candidate is w3 (though an analogous problem arises if we go with w2 instead); hence w1 < w3 (w1 < w2). But w1 can't both be at least as good as w3 is *and* worse than w3 is. (w1 can't both be at least as good as w2 is *and* worse than w2 is.)

In effect, the move to PAIO(c) completely dismantles the person-affecting deontic solution to the nonidentity problem. Not only does it bar what it seems the person-affecting theorist should want to say about the *outcomes* that are to be compared. It also means we must, after all, endorse the result that the *act* a1 is permissible—that is, reject the very result that we earlier said that PAIA* leaves room for: that a1 is wrong.[38]

---

[37] See note 28 above [Pareto principles].

[38] To see, in other terms, why accepting the axiological constraint disrupts our prior solution of the nonidentity problem, consider the following.

If w3 is the outcome that demonstrates that a1 is wrong, then w3 > w1.

But if our interest is in comparing w1 and w2, and if that means we can't know about w3, then we can't know that w3 > w1. Instead we will think that w1 is at least as good as w2 is. (We can't see anything amiss in in w1.)

I.e., we can't see the truth about w1 and w2; we can't see that w1 < w2, that is, that w1 is worse than w2 is.

But if w1 is deemed at least as good as w2 is, and if we have already deemed w2 permissible under our person-affecting theory, then w1 must be permissible as well. But that's an inconsistency, since we know on other (widely accepted, incontrovertible) grounds that w3's being better than w1 is what shows that a1 is wrong.

So: as to outcome comparison, not recognizing w3 means that we can't rank w2 as better than w1. And, as to act evaluation, not recognizing w3 means that a1 and a2 are

In contrast, PAIO* works with PAIA* hand in glove. The necessary condition on w1 being worse than *w2* that PAIO* sets forth is perfectly satisfied: there does or will exist a person p in w1 such that w1 is worse for p *than w3 is*. PAIA* thus opens the door for a further person-affecting principle that deems w1 worse than w2 on the ground that w1 is worse for p *than w3 is*. More generally, we can say just what we want to say about this case: w2 is exactly as good as w3, and w1 is worse than both w2 and w3.

3.5.3 *Unpalatable options for preserving the person-affecting solution to the nonidentity problem*. A theorist completely smitten with the axiological constraint but finding the person-affecting intuition attractive and hence wanting to articulate that intuition carefully might consider one or both of the following options attractive.

(i) One way of preserving our solution to the nonidentity problem while accepting PAIO(c) in place of PAIO* would be to reject the assumption of *transitivity*. Indeed, it might seem that the preceding discussion has shown that that's just what the person-affecting theorist *should* so in order to preserve the person-affecting solution to the nonidentity problem. It might even seem that the preceding discussion shows that to accept the person-affecting approach just is to reject transitivity.[39]

---

both permissible. Hence our arms are tied—we can't solve the nonidentity problem—if we accept the axiological constraint.

[39] The person-affecting intuition and the rejection of transitivity of the *at least as good as* relation between outcomes are often considered to come together. In fact, however, they come together *only if* the person-affecting intuition is formulated by reference to the axiological constraint (which might, e.g., instruct in the context of the mere addition paradox that A+ is at least as good as A) rather than by reference to modally sensitive principles (which will instead leave room for the result that A+ < A).

Moreover, there is good reason to think that the well-formulated person-affecting intuition actually *rescues* transitivity. Thus the person-affecting intuition—related, as it is, to the intuition of narrow neutrality; see part ___ below—rejects the notion that the existence of the additional happy person makes an outcome better. Once we reject that notion, we are in a better position to avoid the repugnant conclusion and still insist on transitivity.

For discussion of the relation between transitivity and alternative forms of consequentialism, see Carter 2015 [in AAA Research].

I would resist that position and indeed that understanding of the preceding discussion. It's a rush to judgment. We should instead first examine whether the telic component of the person-affecting intuition really is subject to the axiological constraint—whether, that is, the classic view really does squeeze the person-affecting intuition into the mold of PAIO(c) and rule the modally sensitive PAIA* out-of-bounds from the start.

(ii) There is a second, and final, option for preserving the deontic person-affecting solution to the nonidentity problem while accepting PAIO(c) in place of PAIO*. We could reject the very tight connection between the evaluation of acts and the evaluation of outcomes. We could then insist that a1 at w1 is wrong even though w1 ≥ w2 ≥ w3.

But that option seems unpalatable as well. For one thing, it seems clear, on grounds that have nothing to do with the nonidentity problem, that PAIO(c) is objectionable. Surely it's not the case that w1 is at least as good as w3 is.

Moreover, if we reject the very tight connection between the evaluation of acts and the evaluation of outcomes, then the questions why we are interested in the latter at all and what we are supposed to look for in the successful candidate immediately arise. I'm not sure, however, that we can answer those questions.

3.5.4 *How the axiological constraint undermines narrow neutrality.* PAIO(c) doesn't just mean that we can't effectively defend the person-affecting intuition against the nonidentity problem. It also means that we must abandon the closely related happy child half of the *procreative asymmetry*—the idea, that is, that it doesn't make things better to bring an additional happy person into existence—and specifically what I will call *narrow neutrality* in Part II.

Suppose that we want to compare w1 against w2. PAIO(c), as we have just seen, would instruct that w1 ≥ w2.

Nonidentity problem aside, on its face that may not *seem* an alarming result. But in fact it is a *quite* alarming result—*if* we want to retain narrow neutrality, according to which w2 (in which p never exists) ≥ w3 (in which p exists and p's

wellbeing is maximized). On that principle, the addition of p in w3 doesn't make w3 better than w2. (Here the principle is plausibly *narrow*; unlike the neutrality intuition itself, that is, what Broome calls the neutral *range* claim, *narrow* neutrality leaves open that the addition of p *in w1 does* make w1 worse than w2). Transitivity, again, tells us that w1 ≥ w3, a result that, as noted just above, seems clearly false. Our only palatable option—if, that is, we keep PAIO(c)—may then seem to be to reject narrow neutrality—to accept, that is, that the existence of the additional happy person *does* make a given outcome better. But we don't want to do that—hence the bind.

3.6 *Evaluating the axiological constraint*

We thus need to get out of the straightjacket that the axiological constraint imposes—not just for the sake of the person-affecting intuition but for the sake of narrow neutrality as well.

But is that really so very hard to do?

3.6.1 *My proposal*.

Suppose our interest is to compare w1 against w2 in a case where w3 exists as an accessible outcome. I want to be able to say that relevant to the evaluation of how w1 compares against *w2* is the fact that w1 makes things worse for p than *w3* does. I want to say that w3's accessibility tells us that w1 is "bad for" p, in the morally relevant sense that it is *worse for* p than is at least some other accessible outcome.

PAIO*, of course, lets us take that fact into account and conclude that the principle's necessary condition on one outcome's being worse than another is satisfied. The necessary condition on w1 being worse than w2 is satisfied, not in virtue of the fact that w1 is worse for an existing or future person p than *w2* is—it isn't—but rather in virtue of the fact that w1 is worse for p than *w3* is.[40]

[40] We're not making the judgment that p is *wronged* in w1; that comes later; that's not what grounds the result that w1 is worse than w2 is. We're just noting a fact not about worseness between outcomes, but about one outcome's being worse *for a person* than another outcome is.

But as we've seen what we would *like* to be able to say runs afoul of the axiological constraint, which itself, as we have also seen, arises out of the effort to squeeze the telic component of the person-affecting intuition into the classic view. The axiological constraint insists that the person-affecting necessary condition on when one outcome is worse than another must be spelled out in a certain way—spelled out, specifically, in a way that approves PAIO(c) but immediately rules PAIO* out-of-bounds.

The question we now face is whether the axiological constraint really is a constraint we have no choice but to accept.

But how we answer that question, in turn, depends on how we answer two others: is the classic view itself a view we have no choice but to accept, and, if it is, does the classic view generate the axiological constraint?

3.6.2 *Inconsistency argument for the classic view*. Arguments for the classic view may seem plentiful. But some of those arguments aren't persuasive and others aren't fully developed.

Thus we noted earlier that on the classic view it is irrational to allow considerations relating to wγ—those *modal details*—to affect—that is, to *change*—how we rank wα against wβ. The basis for that judgment of irrationality was that when the goal is to determine whether an outcome is *better*—and not to determine whether an outcome is *the best*—a correct procedure for making that determination *need* not insist that we look beyond wα and wβ themselves. And if we need *not* look beyond wα and wβ, then it would be irrational to think that looking beyond wα and wβ could change our result. And that may be. But note that from none of that does it *follow* that, for the purpose of comparing wα against wβ, facts about wγ can have *no bearing* on our evaluation. For it remains at least possible that facts about wγ—details that may bear on how wα is to be ranked against wβ—are already reflected, *sotto voce*, in wα and wβ. But if *that's* how the classic view is itself to be understood—as, that is, open to that possibility—then it doesn't generate the axiological constraint.

Other arguments may seem to support a stricter reading of the classic view. Such arguments may cite the axiom of the independence of irrelevant alternatives—for short, the *independence axiom*. According to that principle, how wα compares against wβ is not itself affected by the existence of wγ as a further accessible alternative; if wα is worse than wβ when wγ *is* an accessible outcome in the particular case, then wα is worse than wβ when wγ *isn't* an accessible outcome. It may well seem natural enough—though we shall come back to this point in what follows—to read the principle as ruling out the possibility that facts about wγ—*even if* already reflected, *sotto voce*, in wα and wβ—cannot make a difference to how wα compares against wβ. However, on that reading of the principle, there's very little light between the principle on the one hand and the axiological constraint on the other. So to cite the one in favor of the other isn't persuasive.

Still other arguments might simply appeal to a certain concept of intrinsic value or just to the position—adopted, i.e., by Temkin as critical to the *intrinsic aspect view*—that, it being wα and wβ we aim to compare, we should have no need to look beyond wα and wβ to make that comparison.

I think, however, that the best argument in favor of the classic view, *strictly* construed—construed, that is, so that it *does* support the axiological constraint—is an *inconsistency argument*. Specifically, it's the argument that, without the classic view, strictly construed, we will end up with an inconsistent ranking of outcomes. We'll end up saying that two outcomes *aren't* equally good in a case where the third outcome *is* accessible and that the same two outcomes *are* equally good in a case where that third outcome *isn't* accessible.

We've already introduced a schematic for the three outcome case (Graph 3.5.1). To see how the inconsistency argument works, we need a schematic for the two outcome case as well.

Graph 3.6.2: Two Outcome Case		
wellbeing	w1 (including a1)	w2 (including a2)
+10		
+5	**p**	
+0		*p**

Now, for the three outcome case, we want to say that w2, where p never exists, is equally as good as w3, where p's wellbeing is maximized, and that w1, where p's wellbeing is avoidably *not* maximized, is worse than both. And, as we have seen, we can say just those things under PAIO*.

But for the two outcome case, we want to say that w1 is equally as good as w2 is. And again we can say just that under PAIO*.[41]

But now we have an inconsistency. The modally sensitive PAIO* seems to instruct that both that w1 is equally as good as w2 is and that w1 is worse than w2 is.

The classic view, strictly construed, guards against just such an inconsistency. If w1 is worse than w2 in the three outcome case, then according to the classic view, strictly construed, w1 must be worse than w2 in the two outcome case as well.

I think, however, that this argument to inconsistency fails. Consider, again, any case that fits the three outcome schematic. In any such case, it's clear there will be a causal explanation of w3's accessibility—an explanation rooted in the *modal details* inherent in w1; an explanation rooted in how things, consistent with the laws of nature[42], *could* have been. Specifically: agents in w1 had the power,

[41] The necessary condition is failed, w1 not being worse for p than w2 is, and we infer that it's not the case that w1 is worse than w2 is. And we infer as well that it's not the case that w2 is worse than w1, w2 not being worse for anyone who does or will exist in w2 than w1 is.

[42] And (perhaps) the acts of other agents.

the ability, to bring w3 about and thus make things better for p and they declined to exercise that power. Now consider the case where w3 *isn't* accessible—the two outcome case. There, the reason w3 isn't accessible can also be explained by reference to what is going on in w1. To say that w3 isn't accessible is to say that agents in w1 *lacked* some power, some ability, to make things any better for p than things are in w2.

But that agents have a certain ability in the one world and lack that ability in the other just means that those worlds—w1 in the three outcome case and w1 in the two outcome case—are actually two *distinct* worlds. Worlds, after all, aren't simply *distributions*—bare boned assignments of wellbeing levels to members of a particular population. Rather, worlds come to us with all their details necessarily intact. New details entail new worlds.

And that in turn means that the inconsistency itself is illusory. It's one we immediately avoid upon the introduction of a more exacting vocabulary.

Thus we might say, about the three outcome case, that w1 is worse than w2 is and that w2 is exactly as good as w3 is, and, about the two outcome case, that w1′ is exactly as good as w2′. For the sake of completeness, we can even add that w1′ and w2′ are equally as good as w2 and w3 are, and that w1′ is better than w1 is (that last point, *despite* the fact that the two worlds distribute wellbeing across identical populations in identical ways).[43] These are all perfectly consistent results. No inconsistency or failure of transitivity emerges from anything we have just said.

3.6.3 *Implications for the classic view; independence.* It's worth noting that we haven't here relativized our comparisons to particular cases or choice sets. That means that we remain free to do just what we've done in the foregoing paragraph—complete the ranking, and compare not just the accessible outcomes in

[43] The position that worlds may be distinguished on the basis of their accessible alternatives is not original. [cite] But to my knowledge the point hasn't been discussed as an element critical to any rescue of the person-affecting intuition from the threat posed by the classic view and, specifically, the axiological constraint.

the one case against accessible outcomes *in that one case* but also compare accessible outcomes in one case against accessible outcomes in *another* case.

If the classic view means that, in comparing w1 against w2 in the three outcome case, we need not and indeed *must* not look beyond w1 and w2 and need not and indeed *must* not look at *any facet of w1 or w2 that will tell us whether w3 is an accessible outcome or not*, then we should give up the classic view. But I see no reason why we need to understand the classic view so strictly. Rather, we can instead understand the view to include the idea that we can complete the comparison of w1 against w2 *provided* that we have scrutinized w1 and w2 *closely enough to determine* what the powers and abilities of the agents in w1 and w2 in fact are and, specifically, to determine whether or not p's existence in w1 represents the best agents could have done for p.

In other words, it's the features inherent in w1 and w2—the *modal details* that are part of the very *identities* of w1 and w2—that decide whether w3 exists as an accessible outcome or not. To say that w3 exists as an accessible outcome for the relevant agents is just shorthand—and a convenient and perspicacious shorthand at that—for our saying *something about w1 and w2*. Thus we can after all retain the classic view—*so understood*—but still reject PAIO(c) in favor of PAIO*.

A similar point holds for the independence axiom. We need not understand that principle to say that how we rank w1 against w2 is independent of any facts having to do with w3 *even if* those facts are themselves reflected in w1 and w2. We can instead understand the principle to say that how we rank w1 against w2 *may well* depend in part on facts having to do with w3 *insofar as those facts are reflected in w1 and w2*. Which is just to say that w3 is in very real sense independent to our comparison of w1 against w2, but independent *only* because we find whatever facts about w3 that we need to rank w1 against w2 within the confines of w1 and w2. Looking *carefully* at w1 and w2 is, in other words, precisely the same operation as taking w3 into account.

The upshot? *All that really must go is the axiological constraint itself*, the product of the thought that the only way to squeeze the person-affecting intuition

into the classic view is for the necessary condition on one outcome's being worse than another to be satisfied only if the one outcome is worse for p than the other outcome is *for p*.

3.6.4 *Accessibility axiom*. Residual concerns about inconsistency are addressed by the following principle.

> *Accessibility axiom*. If wβ is accessible to wα, then *necessarily* wβ is accessible to wα.

The accessibility axiom insures that the way of reconciling the results we spelled out for the three outcome case—w1′ is worse than w2′—and for the two outcome case—w1 is equally as good as w2—will hold for any relevantly similar pairs of cases. It guaranties that we won't stumble across still *another* pair of cases where w1′ in the two outcome case turns out to be identical to w1 in the three outcome case. If that third outcome w3 turns out to be *inaccessible* in a given case, in other words, it won't be *w1* that that particular case involves but rather some other outcome altogether.

That a non-standard—that is, person-affecting, or modal—approach relies on the accessibility axiom should not be viewed as itself problematic. After all, the accessibility axiom itself is highly plausible, really perhaps just a product of the concept of accessibility in combination with a very basic understanding of what a possible world consists in.

3.6.5 *Addition plus and the axiological constraint*. Supporting the argument that we should discard the axiological constraint in favor of a more far-reaching, modally sensitive way of comparing outcomes is *addition plus*.[44]

[44] This variation on Parfit's mere addition paradox, in addition to challenging the axiological constraint itself, makes many other points as well. Thus addition plus, like the miserable child half of the asymmetry, nicely shows that some formulations of the person-affecting intuition—e.g., *modal actualism*—will not work. (If w1 is actual, it won't do, e.g., to say that only the actual person p matters morally—that it matters not at all how well off q is in w2 and w3 and hence, e.g., that a1 is wrong or that w1 is worse than w2.)

Graph 3.6.5: Addition Plus			
wellbeing	w1 (including a1)	w2 (including a2)	w3 (including a3)
+10		**p**	
+9	**p**		
+5			**p, q**
+1		**q**	
+0	*q**		

When we compare w1 against w2 without taking w3 into account—without, that is, examining w1 and w2 closely enough to see that reflected in w1 and w2 is the fact that w3 exists as an accessible outcome—my view is that we really have not given ourselves enough information to make the comparison. And if we *do* proceed to make the comparison *without* getting that information—without, that is, coming to understand that w3 exists as an accessible outcome—we may well come to a result that anyone who wants to retain the person-affecting intuition can reasonably reject: that w2 is at least as good as w1 is.[45] To accept that w2 is actually—given w3—worse than w1 is to reject the axiological constraint. To accept the axiological constraint, on the other hand, may well be to give up the person-affecting intuition.

In contrast, consider how the classic view—strictly construed—enables Lazari-Radek and Singer to move without discussion from the result, in a case they have presented as a two outcome case ("natural phenomena . . . cannot be changed

[45] Thus PAIO* allows us to say about this case that w2 is worse than w1 in view of the fact that w2 is worse for q than *w3* is. At the same time PAIO* insists that—in the two outcome case—w2 is at least as good as w1 is. This inconsistency—as noted earlier—we can resolve by distinguishing w1 from w1′ and w2 from w2′.

and . . .compensation [for q] is impossible") that w2 is at least as good as w1, to the result, in a case that they then expand to include a third outcome w3, that w2 *remains* at least as good as w1.[46] But the inference that the classic move—strictly construed—on its face seems to support is questionable. In the two outcome case, the natural phenomena "cannot be changed"; compensation is "impossible"—that's why it is a *two* outcome case. In their three outcome extension, it must be that w2 has now been stripped of those facts; now, phenomena *can* be changed, compensation *is* possible, else it wouldn't be a *three* outcome case.

If, however, we take the scenarios under consideration to be worlds, and understand worlds to have their properties necessarily, then it seems that one case or the other—the two outcome case, or the three outcome case—must be dismissed as impossible.

But there's a more substantive point to be made as well. It seems perfectly consistent and reasonable to say that what we want to say about the first two outcomes is just going to vary, depending on whether the third outcome exists as an accessible outcome or not. And if that point itself is consistent and reasonable, then the door is open to our rejection of the axiological constraint. Having rejected that constraint, we are then in a position to put PAIO* forward as a properly formulated version of the person-affecting intuition, in place of PAIO(c).

3.7 *Remaining work*. Of course, cautions that we should not assume that the person-affecting intuition says one thing—PAIO(c)—when in fact it's better understood to say another—PAIO*—piled on top of cautions that we should be alert to the modal details of our own nonidentity cases doesn't *insure* a solution to the nonidentity problem.

If the nonidentity case doesn't come with any critical modal details—if, that is, our modally sensitive formulations of ontic and telic versions of the person-affecting intuition don't have anything much to see when they do their requisite looking around to take into account all the facts of the relevant case—then the nonidentity problem will remain unsolved. Indeed the inconsistency that the

[46] Lazari-Radek and Singer 2014, pp. 366 and 371.

problem ends with—that a1 both is and isn't permissible—would mean that we must put the person-affecting intuition itself back on the chopping block.

The task now, then, is to show that the most challenging form of the nonidentity problem comes with the relevant modal details. The task now, in other words, is to show that, when the act under scrutiny is clearly wrong, we can explain by reference to those modal details just why the person-affecting intuition avoids the result that the act is permissible.

Chapter 4
Kavka's Pleasure Pill Case

4.1 *Avoiding the mistakes*

Kavka's pleasure pill case is the perfect exemplar of the nonidentity problem in its most challenging form. We are—and, contra Boonin, remain—confident that the act under scrutiny is clearly wrong. At the same time, it's not clear at all to us starting out that that act performed at the particular world has made the child whose plight is so concerning to us to begin with or anyone else who does or will exist in that world any worse off. Rather, it seems to us starting out as though the child surely *owes his or her very existence*—itself an existence worth having—to the act under scrutiny. That means that the necessary condition the person-affecting intuition sets forth, in both its ontic and its telic forms, is immediately failed. The choice is permissible; the one world isn't worse than the other.

To untangle the argument, we must avoid Possible Mistakes A and B. We must present the case in a way that explicitly recognizes its *critical modal details*. And we must apply the *modally sensitive* formulations of both the ontic and the telic components of the person-affecting intuition and avoid the *unduly constricted* formulations. We can then defend the only version of the intuition that we really have any interest in defending to begin with, the version that puts those modal details to work to work in a way that avoids the inconsistency.

4.2 *Modally enriched presentation of the facts*

Thus suppose that an agent, Luc, prior to conceiving a child, pauses to take a *pleasure pill*—a pill that is teratogenic but that produces a mild and transient euphoria in Luc. Luc then proceeds to conceive a child, Andy. Andy's life is clearly worth living. But his life is burdened as a result of the impairment and his overall lifetime wellbeing accordingly reduced.

Now, as Kavka himself notes, had the agent *not* paused to take the pleasure pill—had he, e.g., taken an aspirin instead—the very same child *might* still have

been conceived.[47] But—given the very low probability that any one person would ever have come into existence at all had things been other than just the way they in fact were; given, that is, the *precariousness of existence*—we recognize that the odds are very much against Andy's coming into existence had Luc taken the aspirin rather than the pleasure pill.

Indeed, we make it part of the case that, had Luc not paused to take the pleasure pill, the timing and manner of conception would have been different and Luc would have conceived a distinct child in place of Andy. (Perhaps he rushes to take the pleasure pill and would have taken his time getting to the aspirin.) Let's stipulate further that that distinct child—say, Ruth—would have been better off than Andy in fact is—that is, that Ruth is better off at the closest possible world where Luc *doesn't* take the pill than Andy is at the actual world where Luc *does* take the pill.

Now, this isn't, as we shall see, a *complete* presentation of the pleasure pill case. There's an important probability point we've yet to make. But let's sum up what we have so far:

[47] Kavka 1982, p. 100, n. 15.

Modally Enriched Nonidentity Problem (Incomplete)

Let w1 be the actual world. Let a1 performed at w1 be Luc's act of taking the pleasure pill at w1. Let Andy be a child born seriously impaired at w1 as a result of Luc's taking the pleasure pill at w1.

Let w2 be an alternate accessible world. Luc performs a2 at w2 in place of a1 at w1. Let Ruth be a child nonidentical to Andy born healthy at w2 as a result of Luc's performance of a2 at w2.

We stipulate that w2 is better for Ruth than w1 is for Andy—that is, that Andy has less wellbeing in w1 than Ruth has in w2. We also stipulate that, while Andy's wellbeing is reduced as a result of the impairment, his life at w1 is clearly worth living—that is, that his wellbeing is clearly in the positive range.

Let w3 be a further accessible world in which Luc performs a3 rather than a1—that is, pauses to take the aspirin rather than the pleasure pill—and nonetheless conceives Andy. w3 is better for Andy than w1 is.

Let's recognize that, while w3 is *technically* accessible just prior to performance, the *probability* of Luc's conceiving Andy, given that Luc chooses to take the aspirin rather than the pleasure pill—the probability, that is, of w3 obtaining, given a3—is very low.

Since Andy never exists in w2 and has a life clearly worth living in w1, we infer that it's not the case that w1 is worse for Andy than w2 is. We also stipulate that no one other than Andy who does or will exist in w1 is affected by what the agent does. Hence we infer that it's not the case that w1 is worse for anyone who does or will exist in w1 than w2 is.

We stipulate, finally, the following counterfactual: had Luc not performed a1 at w1, Luc would have performed a2 at w2 and Ruth would have been the one conceived rather than Andy.

[END]

Or, in graph form:

Graph 4.2: Modally Enriched Nonidentity Problem			
wellbeing	w1 (including a1)	w2 (including a2)	w3 (including a3)
+10		**Ruth**	**Andy**
+8	**Andy**		
+0	*Ruth**	*Andy**	*Ruth**

Thus the facts of the pleasure pill case presented in a modally *enriched* framework. Though incomplete, it's a plus that this modal presentation of the facts explicitly recognizes that the future w3 in which Andy both exists and is better off than he is in w1 is accessible to agents. We thus avoid the mistake of ignoring some of the facts of our own case.

4.3 *Objection to modal presentation*

The critic of the person-affecting intuition might object that the details we've added to the case as we work toward completing the presentation unfairly *weaken* the nonidentity problem. The critic might object that we could just as well have added details that instead *strengthen* the problem.

Specifically, rather than completing the presentation by stipulating w3 as an accessible outcome, the critic might object that we could have completed the presentation by *stipulating* that *no such accessible w3 exists*—by *stipulating*, that is, that the outcomes accessible to Luc in the pleasure pill are *exhausted* by just w1 and w2.

Two notes here. (i) One *might* attempt to set w3 aside as *irrelevant* to the evaluation. One *might* have thought such an attempt sensible in virtue of our concession that w3 is highly improbable or our stipulated counterfactual to the effect that, had Luc no performed a1 at w1, he would have performed a2 at w2. We

explored and questioned those rationales earlier. But still it's understandable that one *might* have considered the strategy valid.

In contrast, it seems *undeniable* that w3 exists as an *accessible* outcome—however merely technical; however improbable; however counterfactually doomed—for Luc in the pleasure pill case. A world is *accessible* if it's a possible future the agent has the *ability* to bring about; a world is *accessible provided* it's *not* barred by the laws of nature or by the acts of other agents.

Suppose I need to open a safe to get to a bomb that I can then disarm and which will otherwise blow up the building. But I don't know the combination to the safe. Twirling the dial this way and then that and succeeding in opening the save is nonetheless part of a possible future that is *accessible* to me. I *can* open the safe, even if, very probably, the combination I randomly try *won't* be the right combination.[48]

Ditto the pleasure pill case. There is nothing in Luc's switching from the pleasure pill to the aspirin that would renders it contrary to the laws of nature or acts of other agents for Luc then to conceive Andy. The pleasure pill *isn't* a fertility drug.

This is not to say, of course, that Luc's performing a3 would *guaranty* that w3 would obtain. Rather, including w3 as an accessible outcome in the presentation of the facts of the pleasure pill case simply makes explicit our background understanding that, among all the ways there are out there of Luc's implementing the choice to take the aspirin rather than the pleasure pill, there exists at least one such way—one such act; call it *a3*—that will form part of a chain of acts and events that will yield an outcome—which we call *w3*—in which Andy is conceived and is better off than he is at w1. There is at least one such act, that is, *a3*, that *mimics* a1 in all those spatial-temporal-mechanical features that play any causal role at all in Andy's being conceived rather than, e.g., Ruth, and there is at least one such world in which things then unfold at that world *just as they unfold* in w1.

[48] [Cite for this example.]

The *timing and manner* of Luc's conception at w1 isn't, in other words, *unique* to Luc's performing a1 at w1. The *timing and manner* of Luc's conception *might* be *exactly* as it is in w1 even if a3 replaces a1.

Now, this way of thinking about the pleasure pill case presumes—as noted above—that *the laws of nature are as we understand them to be*. It presumes, in other words, that we come to the table with certain background facts in hand and that those background facts are fair game as we process the case and make the judgments that we then make about the case—including the judgment that a1 is wrong. We understand that not *every* historical blip is actually an essential, or necessary, ingredient to a given person's ever being conceived at all. Plausibly, you couldn't have had genetic parents other than the genetic parents you in fact had.[49] But you could have been conceived in Massachusetts rather than Oklahoma; and Andy could have been conceived had Luc taken the aspirin rather than the pleasure pill. Being conceived in Oklahoma thus was an *accessible* outcome for you. And being conceived Luc's having taken the aspirin rather than the pleasure pill was an *accessible* outcome for Andy. And that's all so, however highly improbable it is that that better-for-Andy accessible outcome would have obtained had Luc not done just as he did.

Can the objection be pressed harder? It's, after all, the critic's hypothetical to do with as he or she pleases. Is it legitimate to present the pleasure pill as just as we have—and then to add the stipulation that after all w3 *isn't* an accessible outcome?

Actually, no—at least, not without a lot of further work. To stipulate that the pleasure pill case is a *two outcome* rather than a *three outcome* case is at odds with our coming to the table with the understanding that the laws of nature are as we understand them to be. It introduces an ambiguity into the case. That ambiguity

[49] One clarification. "Couldn't" here doesn't mean *logically* couldn't, but rather that the technology to make it happen isn't *available* to agents at this point in this world. However, your having genetic parents other than the genetic parents you in fact have would, in addition, seem to be a logical impossibility. I am here assuming—and take as plausible—the genetic origin theory of personal identity.

renders the judgments we make about the case unreliable—whatever our subjective level of confidence might be. And the case thus fails as a counterexample.[50]

The critic can always to a more thorough revision of the case—one that insures that the stipulation that eliminates w3 as an accessible option is *consistent* with the background that we come to the table with, including the presumption that the laws of nature are as we understand them to be; one that insures that we are not trying to have things both ways.

Thus the critic might stipulate, not just that it's a two outcome case, but also that the laws of nature *aren't* as we understand them to be. Thus the critic might have us imagine that Luc takes the teratogenic pleasure pill in a world w1-*alt*. where the pleasure pill acts just like a fertility pill does at the *actual* world. In that far-away world w1-alt., the pleasure pill indeed imposes certain risks on any offspring

---

[50] Cases that effectively counterexample the person-affecting intuition or any other principle cannot smuggle into their facts ambiguities regarding those facts that, for all we know, *destabilize* the judgments that we make about those facts—the very judgments that are meant to show that the particular principle that is being tested is false. But the facts that are included in a given case not limited to what is explicitly said. They include a background of further facts that is itself rooted in our understanding of how the world works—our presumption that the laws of nature are as we understand them to be. Now, we are perfectly free to stipulate that the laws of nature are other than they in fact are. We can make up whatever hypothetical we want, provided just that the hypothetical is itself logically possible and conceptually coherent. What we can't do is present things as though the laws of nature are just as we take them to be—and then stipulate that the laws of nature are after all nothing like we take them to be. I am grateful to Adam Lerner for discussing this general point with me in connection with the trolley cases; [add cite].

Thus, for the sorts of nonidentity cases that I focus on here—the most challenging cases; the cases in which we are confident that the act is *clearly wrong*—we can always stipulate that w3 isn't part of the case. But we do that, we must also make it explicit that it's also part of the case that the laws of nature are other than they in fact are. We can always stipulate that w3 isn't part of the case—but if we do it, we must do it *consistently* throughout the case.

The problem is that as soon as we do that we badly weaken the case. The two outcome form of the nonidentity problem just doesn't challenge the person-affecting intuition as effectively as the three outcome form of the problem does. We can't be as confident that the act under scrutiny is itself clearly wrong. Transforming the pleasure pill into a fertility isn't a minor change in the case. As noted in the text that follows, it, rather, converts the case from one in which the act is clearly wrong to one in which it's not clear at all that the act is wrong.

then conceived. But it's also a pill Luc *must* take if he is ever to conceive any child at all. In that new case, and relative to that new world w1-alt., the world where Luc takes the aspirin instead of the pleasure pill and Andy is conceived *is* inaccessible. But in that new case it's surely at least *unclear* to us that what Luc has done is wrong. Before condemning his choice, we would surely at the very least want to know more about the risks involved.[51]

Thus the act under scrutiny in the new case doesn't meet the *clearly wrong* standard that—I would be the first to concede—the act under scrutiny in the pleasure pill case clearly meets. (Ditto the slave child case, the risky policy case, the depletion case and so on.)

(ii) What makes the nonidentity problem important isn't undermined by our recognizing w3 as an accessible outcome. Nor does that recognition take us very far at all in solving the problem. It's a step, but it's a baby step.

What makes the nonidentity problem such an important problem, rather, happens *after* we acknowledge w3 as an accessible outcome. As the medical examples we discussed earlier suggest, most of us consider the *probabilities* of a given case important to moral evaluation. We consider the calculation of *expected* value of the alternative acts critical to moral evaluation. What makes the nonidentity problem hard, then, is to say how an act can be wrong in virtue of its being bad for a particular person *notwithstanding* the fact that *any alternate act* reduces the probability that *that* person will ever come into existence at all.

We turn to that point now.

[51] This is not to say that *any and all* fertility treatments are permissible. Those involving supernumerary pregnancies are, e.g., notably problematic. But, interestingly, in those cases, for each of the surviving infants, the burden imposed on that infant was not *unavoidable* at all. For each such infant, in other words, there exists an accessible outcome in which that infant was better off. For each such infant, the agents had the ability to avoid the burden on behalf of *that* infant by reducing the pregnancy earlier on by way of selectively aborting some of the *other* developing fetuses.

4.4 *Probability*

As noted, the above presentation of the pleasure pill case, though modally enriched, isn't itself complete. We are still missing some facts, and, in particular, some facts relating to *probability*.

Now, the modal presentation explicitly includes the concession that the probability of Luc's conceiving Andy, given that Luc takes the aspirin rather than the pleasure pill, is very low. Let's just note that there are actually two hurdles that must be overcome in order for the probability of Andy's coming into existence to be anything more than very low, in the case where Luc chooses to take the aspirin. First, having chosen to take the aspirin, Luc must then *implement* that choice *by a3* and not by any of the *many* alternative acts that would equally well implement his choice to take the aspirin. He must, that is, implement his choice by an act that, like a3, mimics a1 in its various spatial-temporal-mechanical respects. And, second, from the point at which the implementing act is itself completed until the point at which conception takes place, the future must unfold in just the way that it does in w1. If, that is, upon the completion of a3, Luc then forbears ejaculation until he returns from a trip around the world—if, e.g., such a *w4* unfolds in place of w3—and, upon the completion of a1, Luc proceeds immediately to intercourse, Andy won't exist.

That's a lot of uncertainty. That uncertainty is built into our modal presentation. But what *isn't* included there is just as important: *that those same hurdles are in place in the case where Luc chooses to take the pleasure pill*. There, too, having chosen to take the pleasure pill, Luc must then implement that choice *by a1* and not by any of the *many* alternative acts that would equally well implement his unfortunate choice. And there, too, having completed the implementing act, the future must unfold in just the way that it does in w1. Luc can't that is, take the pleasure pill and then—departing from the future that in fact unfolds in w1—take the trip around the world and still conceive Andy.

Now, some theorists do not think that the probabilities matter in the context of moral evaluation. Such *actual value* consequentialists will focus strictly on what

Luc could have done (whether, specifically, he could have done more for Andy than he in fact does at w1). Pertinent to their analysis will be *accessibility*.

But for those theorists who—like Kavka—think that the precariousness phenomenon bears on what the person-affecting intuition tells us about the case, *probability* is relevant. What I have argued here, however, is that the precariousness phenomenon cuts both ways. The probabilities involved in the case—whether Luc takes the pleasure pill or the aspirin—are a wash. Expected value theorists, in such a case, will look to the actual value of each of the (equally low-probability) outcomes to determine what the agent ought to do. And in this case that actual value that w1 assigns to Andy is lower than the actual value w3 assigns to Andy.

To complete, then, the presentation, we thus need to include the following addendum:

> Addendum to Modally Enriched Nonidentity Problem
>
> Similarly, while w1 is technically *accessible* just prior the choice, the *probability* of Luc's conceiving Andy, given that Luc chooses to take the pleasure pill, is very low (and is indeed no greater than the probability of Luc's conceiving Andy, given that Luc chooses to take the aspirin.).

4.5 *Modally sensitive principles*

We've moved beyond the modally *impoverished* framework reflected in standard presentations of the nonidentity problem and now situated the case in a modally *enriched* framework.

We've avoided, in other words, Possible Mistake A. To avoid Possible Mistake B is just to make sure that we now apply the *modally sensitive* formulations of the person-affecting intuition in place of the *unduly constricted* formulations of the intuition. Thus we abandon the clearly false counterfactual PAIA(c).

The latter is an especially critical step. It's our rejection of PAIA(c) that allows us to identify the counterfactual stipulation in the modal presentation of our case as the red herring that it is. It's a stipulation of the case—we don't challenge

it; and we recognize that, unlike the two outcome stipulation, the counterfactual stipulation is certainly *not* at odds with our understanding of how the world works. And at first glance it might seem highly relevant to the moral evaluation of a1. But we are now in a position to point out that that stipulation is relevant to the evaluation *only if* we retain the false PAIA(c)—which we, of course, don't want to do.

The revised argument thus will rely not on the false PAIA(c) but rather on the far more plausible PAIA* and PAIA**.

Making that change requires, of course, that we make further conforming changes throughout the argument else we lose validity. The claim must now be, not that a1 at w1 is better for Andy than things otherwise *would* have been, but rather that a1 at w1 is better for Andy than things were under each other alternative act performed at each other accessible world.

Summing up:

Revised Nonidentity Argument/Deontic Form		
Line no.		*Justification*
1	a1 performed at w1 is morally wrong.	Intuition
2′	There exists no alternative act aβ performed at any alternative accessible outcome wβ such that a1 at w1 makes things worse for Andy (or anyone else who does or will exist at w1) than aβ performed at wβ does. [Alternatively: There exists no alternative act aβ performed at any alternative accessible outcome wβ such that the expected wellbeing of a1 at w1 for Andy is less than the expected value of aβ at wβ is for Andy (or anyone else who does or will exist at w1.]	Stipulations
3′	aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and there is an alternative act aβ performed at an alternative accessible outcome wβ such that aα performed at wα makes things *worse for* that person than aβ performed at wβ does. [aα performed at wα is morally wrong *only if* there is at least some person p such that p does or will exist in wα and there is an alternative act aβ performed at an alternative accessible world wβ such that the expected wellbeing of aα performed at wα for p is less than the expected wellbeing of aβ performed at wβ for p.]	PAIA* [PAIA**]
4	It's not the case that a1 performed at w1 is morally wrong.	Lines 2 and 3
5	a1 at w1 both is and isn't wrong	Lines 1 and 4

But this revised argument can be quickly evaluated. Premise 2′, in both its forms, fails. There exists an available act and an accessible world such that that act *both* makes things better for Andy—that is, generates more *actual* wellbeing for Andy—*and* can be expected to make things better for Andy—that is, generates more *expected* wellbeing for Andy—than a1 at w1 does.

Hence the necessary condition set forth in both PAIA* and PAIA** is satisfied, and we never get to the result that a1 at w1 is permissible and thus never face inconsistency.

4.6 *The revised telic argument*

The argument involving the telic component of the nonidentity problem parallels the deontic argument and we won't bother charting it here. What is important to note is that we similarly avoid inconsistency since we never reach the result that it's not the case that w2 is worse than w1—that is, that w1 is at least as good as w2 is. While w2 isn't worse for Andy than *w1* is, w2 is worse for Andy than *w3* is. Accordingly, the necessary condition on outcome worseness that PAIO* sets forth is satisfied. We thus avoid inconsistency—and at the same time leave the door open for other person-affecting principles to instruct that w1 is, after all, worse than w2 is. The upshot? w2 and w3 are equally good—and w1 is worse than both.

Chapter 5

The Two Outcome Form of the Nonidentity Problem

The sort of nonidentity case I have considered here—that is, the three outcome problem—does not itself exhaust the nonidentity problem. We must also take seriously still another sort of nonidentity case—the two outcome problem, where it's simply part of the case that the agent really *can't* do anything more for the burdened child than the agent already has.

An example would be a case in which the agent conceives a child who will be burdened by a genetic condition that we today do not have the ability to cure or substantially mitigate. The child's life nonetheless will be clearly worth living. Moreover—and here it becomes evident that the world we are imagining is quite distant from our own actual world—that child's coming into existence will not make things worse than they are, for any person who does or ever will exist in the one world, in any other accessible world. Bringing the genetically burdened child into existence is, in other words, actually *and* expectationally *maximizing*—for each person who *does or ever will exist* at that world, including the child.

The principles I've ended with—PAIA* and PAIA**—both will deem the act under scrutiny permissible.

But I believe that that act is permissible. At least, it's not even close to meeting the *clearly wrong* standard that the most serious challenges to the person-affecting intuition (contra Boonin) clearly meet.

We *all* come into existence genetically burdened in some way or another; moreover, there's a lot of wrongdoing going around that's connected with procreation even under the best of circumstances. But it's not at all clear to me that, when we really are in a case where, for each and every person, whether existing or future, that person's actual wellbeing and expected wellbeing have been maximized, anything clearly wrong has been done.

Chapter 6

Conclusions

One of my primary aims here has been to bring to the surface certain details—*modal* details—of those nonidentity cases that give rise to what I believe are the most serious objections against the person-affecting intuition—cases, that is, in which the act under scrutiny is *clearly wrong*. I have made my points in connection with Kavka's pleasure pill case. But we can make exactly the same points for Kavka's slave child case, Parfit's depletion and risky policy cases and cases involving historical injustices and environmental failures, including climate change. Those modal details are often left unrecognized or—if noted at all—dismissed as irrelevant to the moral evaluation of the act under scrutiny. But either way a mistake has been made. Those details *are* part of the relevant cases—and *are* relevant to the moral evaluation.

Of course, what we accept as *relevant* to the evaluation is, in the end, a matter of what *principles* we have accepted as properly formulating the person-affecting intuition. But things pedagogically can happen the other way around. Thus the very details of the case—once recognized—can steer us away from problem principles—*unduly constricted* principles—and toward principles that do a better job articulating our underlying intuitions—*modally sensitive* principles.

Thus, whether we are thinking about the seemingly unique ethical issues that arise in *additional* person cases or perfectly ordinary *same* people cases (medical cases; the shoot-Harry-in-the-arm case), focusing on the alternatives beyond simply *what is* and *what would otherwise have been* helps us appreciate that standard formulations of the person-affecting intuition in both its deontic and its telic form will not do. We can understand that a *better* way of capturing what the intuition actually comes to is a less *constricted* way, a more modally sensitive way—a way that explicitly takes into account not just *what is* and *what otherwise would have been*, but *what could have been* as well.

Philosophers may have considered that seemingly obvious way of thinking about the person-affecting intuition to be ruled out-of-bounds in virtue of a certain axiological constraint. I have argued, however, that that constraint surely does not

have the reach they have assigned to it. It does not rule out an interpretation of the person-affecting intuition that requires us to look around and take into account *all* the facts of our own cases prior to evaluating a particular act as wrong or outcome as worse.

Modal Ethics

Part II: Narrow Neutrality

Chapter 7

Intuition and Existence

7.1 *Goals, organization.* [To come.]

7.2 *Terminology*. The terminology for this Part II is mainly the same as for Part I. We'll continue to define the term *person* broadly, connecting personhood with consciousness (including non-human consciousness) and understanding that the person who is *never conscious* at a given world *never exists* at that world.[52] We will continue to distinguish possible *worlds* (*futures*, *outcomes*) from *distributions* and *possible* worlds from *accessible* worlds. The *Accessibility Axiom* will continue to play a role.

How *wellbeing* itself is precisely to be defined will remain, as before, an open question. But we can say that wellbeing indicates how good a person's existence at a given outcome is *for that person*. If a person p has more *wellbeing* in one outcome than p has in another, then the one outcome is better *for p* than the

[52] See Peter Singer. [*Animal Liberation*; *Practical Ethics*.] The term *person* thus includes many nonhuman animals and excludes many human beings. For purposes here, I assume consciousness to be a necessary but not a sufficient condition for a thing's being a person. And I assume that to survive as the same person from one time to another—for the person p at t1 to be numerically identical to the person q at t2—is for consciousness to be knitted together in some fashion or another by a transitive relation of psychological connectedness R. Moreover, I take it that a human or non-human embryo or fetus that hasn't experienced consciousness isn't a person; a human or non-human fetus that has experienced consciousness is in close proximity of, but isn't identical to, a person; and the person that may ultimately develop out of a human or non-human embryo or fetus doesn't come into existence until consciousness emerges. Thus: early abortion involves never bringing a person into existence to begin with whereas late abortion might (depending on facts about when consciousness emerges in humans) involve removing a person from existence.

other. And we will continue sometimes to refer to the well-off person as simply the *happy* person.

Critical for this Part II will be the distinction between *wellbeing* (what is good *for the person*) and the *personal good* (how good the person's existence is *for the world*). If a person p has *more personal good* in one outcome than in another, then (other things equal) it will immediately follow in virtue of the meanings of the terms that that one outcome is *generally* better—is *overall* better, or *morally* better—than the other.[53] Room is thus left for the possibility that a person may have a positive *wellbeing* level in a given outcome even though that person's existence in that outcome contributes *nothing* to the *general* good of that outcome. That is, *wellbeing* may be positive even though *personal good* is zero.

[53] That is: on an additively separable basis. Thus if we say that the existence of a particular person at a personal good level of n at an outcome w1 contributes an amount n to the general good of w1, then we will say as well that the existence of that person at a personal good level of n at an outcome w2 shall also contribute an amount n to the general good of w2.

Chapter 8

Critique of Totalism

8.1 *Totalism*. Traditional consequentialist theories evaluate worlds on the basis of how much of that which makes life precious to the person who lives—how much *wellbeing*—those worlds contain. Such theories are *maximizing* in nature. A world that contains *more* wellbeing is *morally better* than a world that contains *less*.

Traditional consequentialist theories also take it as a given that no one person's wellbeing counts for any more than anyone else's does. Whether it's *your* child, or *my* child, who is assigned an extra unit of wellbeing is immaterial to whether one world, overall, is morally better than another. Such theories are *impartial*—non-agent relative, in other words—in nature.

Relatedly, traditional consequentialist theories are *inclusive* in nature. Your child, my child and everyone else who ever does, will or might exist has full moral status.

Traditional consequentialist theories, finally, make a *very tight connection* between the evaluation of *outcomes* and the evaluation of *acts*, where acts themselves are understood to include *omissions*. If an agent's act creates the *most wellbeing* that that agent can create—where the outcome an act produces is *morally better* than any alternative outcome—the act is *permissible*.[54] Otherwise, it's *wrong*.

Traditional consequentialist theories, so understood, seem to do a good job capturing the *basic maximizing intuition*—that idea, that is, that we ought to do the *best* we can—that is, create the *most wellbeing* that we can—for people. When we fall short of that—when we have created less wellbeing when we could have (other things equal) created more—what we have done is wrong.

The basic maximizing intuition makes sense to us. It seems right. However, the articulation of that intuition here is incomplete. For I haven't said what it is for one act to create *more wellbeing*, or for one outcome to contain *more wellbeing*,

[54] This statement isn't quite right; an agent's *participation* in an act performed by a group may make what the agent has done wrong, even if the agent could not on his or her own have made things better. See Roberts ___ (on nip and collective action problem).

than another. *More wellbeing* might, consistent with what we've said here, mean, for at least some person, more wellbeing *for that person*, or it might mean more wellbeing *in the aggregate*. Even so, the underlying intuition still seems right. But it won't follow that *any old way* of filling in that blank—that *any old theory* that *captures* that intuition but then goes on to say *a lot more as well*—is one we'll be compelled to accept.

Totalism, a paradigm example of a traditional consequentialist theory, fills in the blank by reference to *aggregate* wellbeing. Here, clause (i) sums up totalism's *telic* principle, and clause (ii) its *deontic* principle. Thus:

Totalism:

(i) Where p is any person and W(p, w) is p's individual wellbeing level in a world w,

Total good (w) = $\sum$ W(p, w) for each person p who ever exists in w;

and

wα is *morally better* than wβ iff total good (wα) > total good (wβ);

and

(ii) An act aα performed at a world wα is *obligatory* at a time for an agent iff, for each wβ accessible at that time to that agent such that there exists no accessible wγ that is morally better than wβ, that agent performs aα at wβ.[55]

According, then, to totalism, the agent's *moral obligation* is just to perform that act that we find in the *morally best* of all those worlds available, or *accessible*, to the agent at a given time, where one world is better than another just in case the *total*, or *aggregate*, of all the wellbeing that world creates for each person who does or will exist in that world is greater.

[55] The principle of moral obligation set forth in the text is based on Broome, Toronto talk notes,. For principles governing moral permissibility and conditional obligation, also see Feldman.

In contrast, *person-affecting* views fill in the blank by reference to *individual* wellbeing. We'll defer until later the issue of just how such a view is to be articulated. But—as we shall see—it's very easy to come up with a view that is roughly *person-affecting* in nature but is clearly false.

8.2 *Objections*. I think the articulation of the basic maximizing intuition set forth above starts off well. The question is whether—in the hands of the totalist—it ends well. The connection totalism makes between an act's, or a world's, being *better* and the act's generating, or the world's containing, more wellbeing *in the aggregate* has to give us pause.

For one thing, it's the fact that totalism is aggregative in nature that means that totalism rules out the *happy child* half of the procreative asymmetry—the intuition, that is, that the existence of an additional *happy* child doesn't make a world better and that, other things equal, it's perfectly permissible not to bring that child into existence.[56] Creating more wellbeing *in the aggregate* is something that totalism obligates agents to do. Hence bringing the happy child into existence is something that totalism obligates agents to do. If the focus instead were, for each person, creating more wellbeing *for that person*, and then drawing a distinction between creating more wellbeing by way of bringing that (possible) person into existence and creating more wellbeing by way of making that (existing or future) person better off, the picture would be quite different.[57]

---

[56] This is not to say aggregation *on its own* rules out the happy child half of the asymmetry. Rather, it's to say that aggregation in combination with totalism's *other* features—including its maximizing feature and, relatedly, its unrestricted inclusion, for purposes of evaluating a given world, of the wellbeing level of *all* people, *each and every one* of them, who does or will exist at that world—that rules out the happy child half of the asymmetry.

Moreover, a non-aggregative (or "person-affecting") articulation of the basic maximizing intuition might—depending on its details—itself rule out the happy child half of the asymmetry. Consider, e.g., a theory that draws no distinction between the wellbeing created by way of bringing a happy child into existence and the wellbeing created by way of making an existing or future child better off.

[57] This is not to say there is a moral distinction between existing and future people and merely possible people. All people, in my view, have the same moral status. But it doesn't

The basic maximizing intuition is very strong. But so is the happy child half of the asymmetry. Thus we have to ask: do we have two *intuitions* that are in conflict—the basic maximizing intuition and the happy child half of the asymmetry? If so, that's one strike against intuition. Or, alternatively, does totalism simply offer an imperfect—really a quite *bad*—reading of the basic maximizing intuition? If so, that's one strike against totalism.

Totalism's aggregative feature doesn't just put totalism at odds with the happy child half of the asymmetry. It also means that totalism is at odds with what we think are the right things to say about the repugnant conclusion, replaceability and the infinite population problem. [Brief description of each problem to come.] These Parfitian difficulties are, of course, in addition to a handful of perennial—but just as deep—objections to totalism, including objections based on equality, fairness and priority.

8.3 *Plusses of totalism.* It may seem that we already know quite enough to simply reject totalism—and, more generally, aggregation—outright.

But eliminating *all* forms of aggregation from our moral theory isn't really such a simple matter. Nor is it clearly desirable. Aggregation—or summation, or addition—has its plusses.

For one thing, aggregation may seem conceptually necessary to the basic maximizing idea itself. If *more* wellbeing is morally better than less, doesn't it just follow as a conceptual matter that more wellbeing *in the aggregate* is better than less?[58]

For another, totalism matter of factly, nicely, generates the *miserable* child half of the asymmetry. Totalism thus instructs that, other things equal, the existence of a *miserable* child—the child whose existence is *less* than worth having; the child

follow that all ways of creating additional wellbeing for a given person have the same moral status. See Roberts [asymmetry papers].

[58] I have argued elsewhere that it doesn't. See Roberts 2002.

whose life is *wrongful*—makes an outcome *worse* and bringing the *miserable* child into existence is *wrong*.

That totalism generates the miserable child half of the asymmetry is a reflection of the fact that totalism is aggregative in nature in combination with its being maximizing and inclusive in nature. Those three features together insure that totalism aggregates *without restriction* across the entire population at each and every world subject to comparison—and hence that totalism doesn't cordon off the misery of the miserable child as somehow lacking in moral significance.[59] The misery of the miserable child has full moral significance according to totalism, just as the happiness of the happy child has full moral significance according to totalism. And that's so, *even though* the miserable child's very existence is what is stake; *even though* the miserable child need not ever exist at all; and *even though* the miserable child exists in only one, not both, of the outcomes under scrutiny.

Appreciation of how totalism handles the miserable child half of the asymmetry softens what otherwise might seem a clear affront to intuition. We *thought* the happy child half of the asymmetry was one of our strongly held intuitions. But now we see—that is, we might *seem* to see—that letting go of that intuition is the perfectly reasonable price we must pay to retain the miserable child half of the asymmetry.

Relatedly, totalism generates *stable* results. Suppose that in the end the choice is made *not* to bring the miserable child into existence. That fact doesn't,

---

[59] If the theory cordoned off the misery of the miserable child as somehow lacking in moral significance, then theory couldn't also be *fully maximizing* in nature: one world could turn out to be better than the other merely as a function of its having ignored the plight of the miserable child.

More generally, it might seem that we could preserve the happy child half of the asymmetry while retaining aggregation by simply restricting the scope of those individual persons whose wellbeing levels matter for purposes of aggregation. It might seem, e.g., that we could simply say that the happy child's wellbeing level is outside the scope of the aggregative function in virtue of the fact that that child exists in one but not the other of the two worlds that are the subject of our comparison. But it's widely recognized at this point—by Singer; Arrhenius; and others—that that restriction fails. As Singer argues, that sort of approach would compel us to reject the miserable half of the asymmetry, something we are loathe to do.

according to totalism, somehow render the *nonactual* world better than it would have been had it been *actual*, or the unperformed choice to bring that child into existence *permissible*. The misery of the miserable child has moral significance, according to totalism, whether that child ever exists or not. Consequently, our moral evaluation of the act of bringing the miserable child into existence does not shift depending on whether that act happens in the end to have been performed or not.[60]

And there is more. The fact that totalism is aggregative in nature comes with many theoretical advantages. [[[It allows for easy theoretical check for transitivity; convenience of pairwise comparison.]]]

8.4 *Can we retain the happy child half of the asymmetry?* It might seem that any theory that displays the virtues we've just associated with totalism—and specifically with the fact that totalism is maximizing, aggregative and inclusive in nature—will immediately rule out the happy child half of the asymmetry.

John Broome, however, explores whether that's in fact so. Though he doesn't put things (or perhaps even, perhaps, conceive things) in this way, his discussion of what he calls the *neutrality intuition* can be viewed as a discussion of whether totalism can be corrected in a way that preserves the happy child half of the asymmetry *without* abandoning aggregation.

Thus Broome argues that the neutrality intuition leads to inconsistency. But—as we shall see—his argument against the intuition makes no reference to aggregation. Rather, the principles that Broome relies on to show inconsistency are, instead, principles that find wide acceptance by aggregationists and non-aggregationists alike.

My point then is just that *even the aggregationist* might have an interest in whether the neutrality intuition can be made to work—should such a theorist want to retain the happy child half of the asymmetry *without* jettisoning aggregation. The point is worth mentioning here since—as we shall also see—we might well want to locate ourselves in just that camp.

---

[60] Hence no violation of Rabinowicz's Principle of Normative Invariance.

But for the moment to the topic is Broome. Many philosophers have found his inconsistency argument against the neutrality intuition compelling. I, too, find it compelling—at least given a certain restriction. I think it clearly shows that the neutrality intuition—in the form of what we shall call the *neutral range* claim, and subject to the same restriction—must go. But we can—and will—nonetheless question is whether the happy child half of the asymmetry must go as well.

Chapter 9

Correction to Totalism

9.1 *The neutral range claim*. We can think of the *neutrality intuition* as an attempt to correct a defect in totalism—or at least as an attempt to take seriously *other* philosophers' concerns that totalism is in need of correction.[61]

As Broome articulates it, the neutrality intuition states that there is a *neutral range* of existence such that a person's existing within that range in a given outcome does not, other things equal, make that outcome morally better or worse but is rather *neutral* in its effect. The neutrality intuition, as Broome articulates it, is really just what we can call the *neutral range claim*.

Thus he writes that "for a wide range of lives the child might live, having a child seems an ethically neutral matter."[62] And: "[T]here is some range of wellbeings (called 'the neutral range') such that, if the extra person's wellbeing is within this range, the two distributions are equally good," where the term range is meant to indicate "more than one member; the idea applied for several different levels of wellbeing."[63]

Broome restricts the neutrality intuition—that is, the *neutral range claim*—to cases in which the child's existence falls into the "neutral range." I shall assume that that range is meant to include the very cases of interest to us in the context of our consideration of the happy child half of the asymmetry—that is, cases in which the child's existence is *unambiguously* worth having; cases in which the child's existence *isn't* marginal and *isn't* just *barely* worth having.

On that assumption, the neutral range claim easily generates the result that, other things equal, the outcome in which the happy child exists *isn't morally better* than the outcome in which the child never exists.

---

[61] [Thus Broome cites Narveson. I, too, cling to the happy child half of the asymmetry, as does Heyd.]

[62] Broome 2004, p. 144.

[63] Broome 2004, p. 146.

Though Broome sometimes makes reference to the evaluation of *acts*—when he notes, for example, that there seem to be cases in which "having a child seems an ethically neutral matter"—he doesn't as a matter of theory accept the very tight connection between the evaluation of outcomes and acts. But if we do find that connection plausible—and I do; I am otherwise unclear what the purpose of determining whether one outcome is *morally better*, or even *generally*, or *overall*, better, than another is—we can see that the neutral range claim also generates the result that, other things equal, the agent's not bringing the happy child into existence is *perfectly permissible*.

Restricting the scope of the neutral range claim to existences that fall into the neutral range enables that claim to support the happy child half of the asymmetry *without* denying the miserable child half.[64] Thus we can say that the existence of the miserable child—the existence that is *less* than worth having—falls *outside* the neutral range. As such, the existence of the *miserable* child isn't caught by the neutrality intuition. And we thus are free, on other grounds, to say that the existence of the miserable child makes the outcome *worse* and that the agent's bringing that child into existence is *wrong*.

Just then to note: nothing we have said so far about the neutral range claim requires us to *abandon aggregation* in order to accept the intuition. What we must do, instead, is abandon a *fully inclusive* commitment to maximization.[65]

Broome notes that he himself finds the neutral range claim attractive. But he argues that in the end we must reject it as inconsistent.

9.2 *Narrow neutrality.* I find Broome's argument against the neutral range claim compelling. But I will argue in what follows that we can accept Broome's argument but at the same time recognize an intuition *behind* the intuition—an intuition I will call *narrow neutrality*. I will, in other words, argue that Broome's

[64] Thus it isn't *coming into existence* per se that we set aside (as Heyd sometimes seems to do) as having no effect on how good, overall, an outcome is. Rather it's the coming into existence *within a certain range* that has no effective.

[65] See note [8] above [fully maximizing in nature].

argument against the neutral range claim gives us *no reason at all* to think that we cannot accept narrow neutrality if that in the end is what we want to do.

But I have a further purpose as well. I will argue that we can retain narrow neutrality *without* setting aside all semblance of the aggregative function that, alongside maximization and inclusion, are responsible for the many plusses that come with totalism.

Specifically, I will argue that we can retain narrow neutrality consistent with the very principle that Broome himself deploys in order to rescue an additive approach from a number of traditional objections against totalism, including those based on the values of *equality*, *fairness* and, perhaps, *priority*. Thus Broome argues that his own formulation of Harsanyi's theorem—what I will call *P** in what follows—avoids just such objections. My argument will be that there is no reason to think that P* cannot be understood to avoid objections based on our *existential* values as well. If the *egalitarian*, or the *prioritarian*, can accept P*, so can the *narrow neutralist*, that is, the *existentialist*, accept P*.[66]

Now, it may seem that the additive P*—and, more generally, that any additive, or aggregative, approach—will force on us all the things that concern us about totalism—all the things, that is, that we *don't* like about traditional forms of consequentialism. P* is undeniably aggregative in nature: it aggregates by way of simple summation across the population of a given outcome to determine whether that outcome is overall (Broome says *generally*; I would say *morally*) better or worse than another. Thus, just like totalism, P* may seem not to "take seriously the distinctions between people."[67] And it may thus seem immediately to rule out

[66] Broome presents P* as his own interpretation of Harsanyi's theorem. See GPG. Others refer to it as *additive separability*. The idea is that the good each additional person's existence contributes to the overall good of the outcome is independent of facts about the good other people's existences contribute to the overall good of that outcome. Specifically, the good the additional person's existence contributes to a given outcome is not deflated by the fact that the average good existence contributed by others who do or will exist at that outcome is higher, nor is it deflated by the fact that the number of well-off people who do or will exist at that outcome exceeds a certain level.

[67] Rawls, *A Theory of Justice*.

considerations of equality, fairness and priority—right along with the happy child half of the asymmetry.

In fact, however, P* isn't so closely tied to totalism. Where totalism deploys the unadorned concept of *wellbeing*, P* instead puts the *highly* adorned concept of the *personal good* to work. It's that fact, Broome argues, that turns P* into a far more defensible principle—a principle capable of recognizing a myriad of values that totalism itself is completely oblivious to.

At the same time, in part as a function of the fact that it is additive in nature, P* has the very plusses we earlier attributed to totalism—the plusses, that is, but not the minuses. [It thus accommodates the miserable child half of the asymmetry; it's a straightforward way of articulating the basic maximizing intuition; it's inclusive; it's results are stable; it allows for pair-wise comparisons between outcomes and it helps us check our work—check for, e.g., failures of transitivity.]

One note. As we go about fitting narrow neutrality into a framework that includes P* but avoids Broome's inconsistency argument, we shall discover that an *inversion* of the calculation of the *personal good* from that which Broome himself may have had in mind in order. Inversion will be critical to understanding just how P* itself can account for narrow neutrality. But inversion will also help us explain the deeply held intuitions we have in connection with some of the other many problem cases in population ethics as well. Or so I will argue in what follows.

Chapter 10

The Neutral Range Claim

10.1 *Tradeoff.* We've already noted that totalism is at odds with the happy child half of the procreative asymmetry. The neutral range claim tries to correct totalism in a way that preserves the happy child half of the asymmetry without forcing us to reject the miserable child half of the asymmetry.

For both halves of the asymmetry, it's part of the case that the existence of the additional child affects no one else. But the defect in totalism that is at play in its treatment of the asymmetry comes to the surface even more clearly, I think, in cases in which the existence of the additional child *does* affect others. So let's start by noting how totalism fails in a case of that sort—I'll call it the *tradeoff case*—and how the neutral range claim seems initially to help.

The tradeoff case involves just two options: bringing a happy person into existence by way of imposing a steep decline in wellbeing for a distinct person and avoiding that steep decline in wellbeing on behalf of that distinct person by way of leaving the happy person out of existence altogether. It's immaterial to the case whether that distinct person is an *already-existing*, or a *future*, person.

The outcomes displayed in Graph __ (i) exist as *accessible* outcomes in the case and (ii) *exhaust* those outcomes. Bold face means the indicated person does or will exist in the indicated outcome, and italics paired with the "*" means the indicated person never exists in indicated outcome.

Graph 10.1: Tradeoff	Wellbeing	w1	w2
Life well worth living	+10	**George**	**Jill**
	+9 … +2		
Life barely worth living	+1		**George**
	+0	*Jill**	

We are to suppose here that George's life at +10 goes really well for him in w1 and he is considerably worse off, at +1, in w2. If and only if George's wellbeing is reduced from +10 to +1, Jill will exist and have a life in w2 that at +11 is a little better than George's life is in w1. Total wellbeing being greater in w2 than in w1, totalism immediately implies that w2 is morally better than w1 is—and that it would be wrong to protect George at Jill's expense.

But both the telic and the deontic results here seem false. If we agree that that's so—and my aim here is not to argue that it is but rather to query whether we can consistently take the position that it is within a framework that otherwise seems plausible to us—then we will consider the tradeoff case to represent a serious problem for the totalist.

10.2 *How the neutral range claim helps.* Totalism implies that the two ways of adding wellbeing represented in the tradeoff case—adding wellbeing to Jill's stock and adding wellbeing to George's—work equally well.

The neutrality intuition—in the form of the neutral range claim—comes along and says that that's a mistake. It's a mistake to see Jill's wellbeing in w2 as adding to the total good of w2. Rather, Jill's existence in w2, despite her relatively high wellbeing level, should be counted as morally *neutral*—as an addition that

doesn't make the outcome better or worse.[68] We understand, on other grounds, that the effect on George of bringing Jill into existence isn't *neutral* at all—that what is done to George in w2 *does* make w2 worse. And we can then see how a perfectly routine account of the case would proceed to get us to the results that w2 is *overall* worse than, not better than, w1 and that bringing Jill into existence at George's expense is wrong.[69]

Those results seem entirely plausible. When the tradeoff is between bringing one person into existence and avoiding a loss on behalf of an existing or future person, it's the latter, not the former, that makes things better.

10.3 *Broome's inconsistency argument*. The neutral range claim seems to offer just the sort of intuitive correction totalism needs. Broome argues, however, that the neutral range claim is inconsistent. Consider the following *three outcome case*. As before, we stipulate that the displayed outcomes all three exist as *accessible* outcomes within the particular case. I should go ahead and note that that restriction—the *intra-case* restriction, I'll call it—is one that Broome himself disputes. Nonetheless, we'll first work through the argument with that restriction in place. For it's that form of the argument that—I believe—tells us something important about the neutrality intuition. It tells us the neutrality intuition, in the form of the *neutral range claim* and subject to that restriction, is inconsistent—and it suggests a better way of articulating the underlying intuition—the intuition *behind* the intuition, that is, *narrow neutrality*. We'll then consider how the argument unfolds *without* the restriction in place. There, I'll make the case that the argument fails.

[68] All we need to assume here is that Jill's wellbeing in w2 falls into the neutral range; that George's wellbeing in w2 might fall *below* that range is incidental to how the neutrality intuition applies since it isn't George's existence that is at stake.

[69] I won't delay things by laying out the specific principles here but I think they are both obvious and highly plausible. But see Roberts [Abortion and the MS of MPP].

Graph 10.3: Three Outcome Case	Wellbeing	w1	w2	w3
	+10			**Paula**
	+5		**Paula**	
	+0	*Paula**		

It's an assumption of the case that Paula's existence in both w2 and w3 falls into the neutral range. Let's suppose that she has a very good life in w2 and an even better life in w3. We then compare w2 against w1. According to the neutral range claim, Paula's existence in w2 is neutral—it doesn't make w2 either better than or worse than w1. It follows, given the simplicity of the case, that w2 and w1 are equally good.[70] The neutral range claim produces parallel results when we turn to compare w3 against w1: w3 and w1 are equally good. Assuming that the *equally as good as* relation between outcomes is both transitive and symmetrical, we then infer that w2 is equally as good as w3. But we understand, on other grounds, that w2 is worse than w3 is. Given that w2 and w3 both exist as accessible outcomes per the intra-case restriction, the principle that moves the argument forward can be understood to be a simple, straightforward Pareto-like principle: where two such accessible outcomes contain exactly the same people, and one outcome is better for at least one person and worse for none than the other outcome is, then the one outcome is itself worse. Hence we have an inconsistency. w3 can't be equally as good as *and* better than w2 is.[71]

---

[70] I don't see the cases at issue in either Part I or Part II as challenging the completeness of the betterness relation. If X isn't better than Y and Y isn't better than X, then X is equally as good as Y is. I leave aside the question whether more complicated cases may represent legitimate challenges to completeness. [Cite R. Chang.]

[71] Broome 2014, pp. 146-147. Broome formulates the inconsistency argument not in terms of *wellbeing* but rather in terms of what he calls *wellbeing*. Now, his use of the term

Conceding both the transitivity and symmetry of the *equally as good as* relation and the claim that w2 is worse than w3 is, to avoid inconsistency we are forced to reject one or both of our two neutrality results. We must then reject either the result that w2 and w1 are equally good or the result that w3 and w1 are equally good. To reject either one or both those results is itself, of course, to reject the neutral range claim. So we reject the neutral range claim.

Notably, however, what we *can't* validly derive from Broome's argument is that *both* disjuncts are false. The inconsistency shows we must reject *one or the other* of the two disjuncts. But it doesn't show that we must reject *both*.

For example, we can avoid inconsistency by claiming that w2 and w1 *aren't* equally good. Specifically we can say that w2 is *worse* than w1. We are then free to insist that w3 and w1 *are* equally good, that is, that Paula's existence in w3 as compared against w1 *is* morally neutral.

Broome's argument is thus *not*—at least not *immediately*, on its face, without further supplement—an argument against the claim that there is *some* level of wellbeing such that Paula can be brought into existence at that wellbeing level

---

wellbeing in 2004 is at least at some points arguably synonymous with what he in 2015 calls the *personal good*. As we shall see, the *personal good* is itself an amazingly accommodating, *highly* adorned concept. The problem is that if his inconsistency argument is meant to make use of that *latter* concept—that is, the concept expressed by *wellbeing* in 2015— then we can't even set up the case for purposes of testing the neutral range claim without tripping over our own terms. Thus, *by definition*, the existence of a person in an outcome at a positive level of the *personal* good *increases* the *general* good. Broome 2015. But that means that if, in the 2004 inconsistency argument, we take +10 in w3 refers not to wellbeing but rather to the personal good, then Paula's existence at +10 in w3 would *immediately* generate the result that w3 is better than w1, a result that would in turn *automatically* rule out the neutral range claim. Broome would then have no need to draw on the simple, straightforward Pareto-like principle to show inconsistency. But he clearly *does* draw on *some* version of that very principle. (Now, just *which* version he means to draw on will be up for discussion later. See part ___ below. But for now the important point is that he draws on *some* such principle.) Hence it seems we should understand the inconsistency argument to refer not to the *personal good* but to the unadorned *wellbeing* instead.

I suspect that is indeed just how Broome means *wellbeing* to be construed in this particular context. And that seems so, despite the fact that things are further confused by Broome's own 2004 name for the Pareto-like principle: the principle of personal good. Broome 2004, p. 120.

in that outcome without making that outcome either better or worse than w1 is. Rather, it's an argument against the claim that there exist *two or more* such levels.

At points, Broome himself restricts the conclusion he draws from the inconsistency argument in exactly that way. Thus he says that the argument tells us that there exists *at most* a *single neutral level*, a "sharp boundary," in the three outcome case—at most a *single* level of wellbeing such that bringing Paula into existence at that level does not make things morally better or worse.[72]

But that result conforms *perfectly* to what we might well consider the intuition *behind* the neutrality intuition—the intuition I will call *narrow neutrality*. We can thus *easily* let go of the idea—derived from the neutral *range* claim—that bringing Paula into an existence that makes things *worse* for her when things could have been *better* is morally *neutral*. We can quite happily instead say that Paula's existing at a *avoidably lesser existence* in w2 makes w2 *worse* than w1 is. We can quite happily instead say that, given w3, w2 is worse than w1 is, which is just to say that w3 shows that w2 is worse than w1 is.

But the moment we agree that w2 is worse than w1 we avoid the inconsistency while giving ourselves the option of retaining narrow neutrality. We thus can say that, in the particular case and for the particular person, Paula, it's in w3, not w2, where the "sharp boundary" of the neutral level is itself achieved. According to *narrow neutrality*, it's at that level and at that level alone that Paula's existence is morally neutral.

We can thus say about the three outcome case exactly what I think we *want* to say. Though Paula's existence in w2 makes w2 *worse* than w1 is, her existence in w3 *doesn't* make w3 *better* than w1 is.

A critical point. None of what I have said so far indicates, for purposes of understanding narrow neutrality, how the *neutral level* is itself to be defined. We can, however, note that it's not plausible to say that the neutral level is, e.g., always +10. If our facts were just slightly different—if Paula's wellbeing in w3 is not +10 but rather +9—we would still want to say that her existence in w2 makes w2 worse than w1 but her existence in w3 is neutral. Or if the case includes still a fourth

[72] Broome 2004, p. 142.

accessible outcome—a w4 just like w3 except that Paula's wellbeing in w4 is +11 rather than +10—we would then want to say that her existence in w3 isn't after all neutral. We would want to say, that is, that w4 shows that Paula's existence in w3 makes w3 worse than w1. Thus what counts as the neutral level will be *case-*, or *context-, dependent*. Nor, for reasons having to do with cases in which what is at stake is the existence of two or more people and the changes between one outcome and the other constitute *merely reversing changes* (Vallentyne), do we want to say that the neutral level is the *maximal* level wellbeing that might be achieved for a given person within a given case. So there is no simple formula for calculating the neutral level for a given person in a given case. We shall thus need to come back to this question. But that the neutral level isn't rigidly fixed for all people and for all cases doesn't mean that it doesn't exist.

10.4 *Interpreting the argument*. Let's go back to the original three outcome case and Broome's inconsistency argument. Would Broome concede that that argument against the neutral range claim *doesn't* rule out the position that w3 and w1 are equally good? That it doesn't, that is, rule out *narrow* neutrality? It seems that he surely would have acknowledged that his argument opens the door to the position w3 and w1 are equally good if he thought that it did. Moreover, there is some reason to think that Broome might have meant for his inconsistency argument to rule out from the start the position that w3 and w1 are equally good. We consider both sides of the question here.

Consider how Broome introduces the neutrality intuition.

> *Neutrality intuition*: "Adding a person to the world is very often ethically neutral."[73] And, quoting Narveson, "'we are . . . neutral about making happy people.'"[74]

Immediately we have a question. "Often"? What's "often"?

[73] Broome 2004, p. 143.

[74] Broome, Stern Report contribution, p. 17.

The locution "often" might be meant just to recognize the *miserable child* exception to the neutrality intuition—to note, that is, that cases where the person's existence falls *below* the *neutral range* are outside the intuition.

But it's also possible that Broome's "often" is meant to recognize exceptions *beyond* the miserable child exception. It's possible, that is, that Broome's "often" is also meant to recognize an exception to neutrality Narveson himself would likely approve—that is, the *avoidably lesser existence* exception. If so, then Paula's existence at w2 is (like the miserable child's existence) would fall outside the intuition.

On this reading, the neutrality intuition *doesn't* imply that Paula's existence at w2 is neutral but rather implies just that existence is neutral often *enough*—*enough* being in the case at hand just *once*, that is, Paula's existence in w3—to instruct that w3 isn't morally better than w1 is.

That would mean, in turn, that the inconsistency argument itself targets, not the claim that w2 is equally as good as w1 *and* w3 is equally as good as w1, but rather just the claim that w3 is equally as good as w1 is.

That reading of Broome may seem at odds with what he says about the neutrality intuition. Thus he explicitly notes that he uses the term *range* to "imply . . . more than one member."[75] But consistent with that point we might say that Paula's existence in w3 falls within the neutral range *as does her existence at various other outcomes in various other cases* but that her existence in w2 falls below it.

The idea that Broome meant his argument to rule out the position that w3 and w1 are equally good might also seem at odds with his presentation of the inconsistency argument itself—its simplicity, its elegance, the seemingly obvious principles (transitivity, symmetry, the seemingly straightforward Pareto-like principle) that moved the argument forward. If the intuition Broome meant to prove inconsistent was *narrow* from the start, then the argument he would have needed to

[75] Broome 2004, p. 146.

launch would have been considerably more complicated than the argument that he in fact describes.[76]

But consistent with that point perhaps Broome's statement of his own argument is itself just a sketch. Perhaps he takes for granted we'll fill in the gaps ourselves.

These two points together suggest we may have a *little* room to interpret his argument as targeting the claim that w3 and w1 are equally good—as, in effect, targeting what I am calling *narrow neutrality* here.

But there's still a third point in favor of that idea. It's *where Broome goes* once he's completed the inconsistency argument itself. Thus let's call the conclusion he reaches in the inconsistency argument—whatever the content of that conclusion—the *intermediate* conclusion. Broome then at various points seems to draw a further conclusion, an *ultimate* conclusion, to the effect that w3 is *better* than w1 is. That ultimate conclusion would indeed seem to follow—we'll see why in the next paragraph—*if* the intermediate conclusion itself is that it's not the case that w3 and w1 are equally good—if, that is, the intermediate conclusion itself is just that *narrow* neutrality is false. But if all Broome has to work with there is that the neutral *range* claim is false, then that *ultimate* conclusion *doesn't* follow at all. It remains pie in the sky.

How would that intermediate conclusion, that w3 and w1 aren't equally good, help Broome get to his ultimate conclusion, that w3 is better than w1? Well, we really don't think that Paula's existence at w3 makes w3 *worse* than w1 is (we

[76] It would have been an *iterative* argument, one that would have involved the claim that just as w3 shows that Paula's existence at w2 isn't neutral, so does an outcome w4, where w4 is just like w3 except that w4 is better for Paula than w3 is, show that Paula's existence at w3 isn't neutral, and so on. And it would have been an argument that relies, not on the simple, straightforward Pareto-like principle that instructs that, in a case like the three option case, where w3 exists as an accessible outcome, w3 is better than w2, but rather a more contestable principle, one that asserts that w3 has a deflationary effect on the value of w2, making w2 worse than w1 even in the case where w3 *doesn't* exist as an accessible outcome to w2. We return to this question in part ___ below. But the upshot would be that the seemingly obvious proposition that w3 is better than w2 isn't really obvious at all if we stipulate from the start, not that w2 and w3 are accessible outcomes within the same case, but rather that the w2 we are talking about may hale from a different case altogether, one in which w3 *does not exist* as an accessible alternative.

are not, after all, Benatarians). So let's—for the moment—take it as an assumption that that's so.[77] But if it's not the case w1 and w3 are equally good and it's not the case that w3 is worse than w1, then we are left to conclude that w3 must, after all, be *better* than w1 is—that Paula's existence in w3 must, after all, make things morally *better*.[78] In short: we should *agree* that, if it's not the case that w3 and w1 are equally good, then w3 is better than w1 is.

But that's a very strong ultimate conclusion, a conclusion with profound practical implications. Broome thus writes that "If [the neutrality intuition] were correct, it would give us a quick answer to the question about the value of extinction: it is neither good nor bad. *But actually the intuition is false*."[79] And since no one really thinks that, other things equal, the *non*-extinction of the species—the *survival* of the species—would make things *worse*—at least, so we shall assume for purposes here[80]—we are left to conclude that it would make things *better*. And: "*Given that the neutrality intuition is false*, the extinction of humanity might be a very great disaster indeed. It would prevent the existence of huge numbers of future people, and the existence of each one of them might well have been a good thing."[81]

Hence the question of interpretation. Is there more to the simple inconsistency argument itself than what we have so far seen? Should we understand

---

[77] That w3 isn't worse than w1 really is just an assumption. It's entirely plausible given how we have here understood Broome's argument. However, on a reconstruction of that argument that we will consider later on, it's an assumption we shall need to question. See part [__] below.

[78] Again—contra Chang—I am taking for granted that in this simple case issues of comparability (or commensurability) do not arise. w1 and w3 are equally good, or w1 is worse than w3, or w3 is worse than w1. See note __ above.

[79] Toronto climate change remarks p. 8 (emphasis added). And "If the intuition of neutrality is correct, the extinction of humanity will be much less of a catastrophe than it might seem at first. . . . Actually, the intuition of neutrality has to be false. It cannot be consistently fitted into any theory of value." Broome, Stern Report contribution, p. 17.

[80] See note [28] above.

[81] Broome, Stern Report contribution, p. 17 (emphasis added).

it to purport to show, not just that it can't be that *both* w2 and w1 are equally good *and* that w3 and w1 are equally good, but rather that it can't be that w3 and w1 are equally good and hence, we agree, must be that w3 is better than w1? Is what seemed to be a simple inconsistency argument not really so simple after all?

Our purpose in life does not, of course, lie in interpreting Broome. Our purpose rather is to identify and then evaluate problems that might arise for *narrow neutrality*.

We accordingly face some worrisome possibilities. The first is that the simple argument isn't so simple after all—that it doesn't open the door to narrow neutrality and the position that w3 and w1 are equally good but rather annihilates narrow neutrality along with the position that w3 and w1 are equally good. The second is that we have gotten the inconsistency argument itself right but that there's a further argument that builds on that argument and that itself shows that it's not the case that w3 and w1 are equally good.

Whatever we find, the pressure to inquire can't be ignored. For it seems clear that Broome, somehow, thinks we can get to the result, not just that the neutral range claim is false, but that w3 is better than w1. It seems clear that on his view narrow neutrality is false. We need to understand just why that is so.

11.4 *Summing up*. (i) As conventionally formulated, traditional consequentialist theories—for example, *totalism*—imply that, other things equal, adding a person whose wellbeing level is positive makes a positive contribution to the total good of the world. We've noted totalism faces many problems.

(ii) The *neutrality* intuition comes along and claims that such contributions are *often* not positive but rather *neutral*. Broome's simple inconsistency argument—as original presented—effectively shows that that intuition—understood as the neutral *range* claim—is false. Consistent with that result, however, we can nonetheless accept *narrow* neutrality, w1 and w3 are equally good—that is, that Paula's existence in w3 is indeed neutral—but that w2 is worse than w1—that is, that Paula's existence in w2 makes things worse.

(iii) We now face two alternate possibilities:

> (iii.a) Per a further, not-yet-identified argument that we accept in place of the simple inconsistency argument or in addition to the simple inconsistency argument, we will be forced to reject that last claim—forced, that is, to reject not just the neutral *range* claim but also *narrow neutrality*; or
> (iii.b) We won't identify any such further argument, or will identify *and* reject it, and thus be left with the room we need to *retain* narrow neutrality.

A full investigation of our options here requires us to understand a bit more about Broome's overall framework, including his construction of Harsanyi's principle, that is, P*. We turn to that work now.

Chapter 11

Additivity

11.1 *Broome's additive framework.* We can identify (at least) two further arguments that target *narrow* neutrality and aim to force the result that w3 and w1 aren't equally good and thus—on the assumption that w3 isn't *worse* than w1[82]—the result that w3 is *better* than w1. But a good understanding of how and whether those arguments work requires reference to aspects of Broome's work that go beyond his inconsistency argument.

We might, in any case, be interested in exploring Broome's overall framework for reasons that don't immediately relate to the procreative asymmetry or narrow neutrality. Additive in nature, Broome's overall framework comes with many of the plusses we earlier attributed to totalism. Yet by its very design it's meant to avoid some of the standard objections against totalism, including objections based on equality, fairness and, perhaps, priority.

Thus the first order of business in this Chapter 11 is to describe Broome's additive framework. Second, we identify two further arguments against narrow neutrality. And then third: we reject those arguments.

11.1 *Additivity, the personal good and the general good.* We've seen that Broome explores—and rejects—the correction to totalism proposed by the neutral range claim. But it isn't just our *existential* values that totalism—or indeed *any* view that calculates the total good of a world via a simple summation across individual wellbeing levels—seems impervious to. *Other* values that also seem left out of the picture by totalism but that, Broome concedes, a plausible theory may well need to recognize include values of *fairness*, *equality* and, perhaps, *priority*.

When Broome seeks a correction to totalism that is itself additive in nature, it's those *other* values, *not* our existential values, that he aims to show that additivity can accommodate. Let's see how the reconstruction works in connection with those other values. We'll then ask the question whether that same reconstruction can be extended to cover our existential values as well.

---

[82] See note 28 above.

Two steps are critical to Broome's reconstruction. The first is the concept of the *personal good*. The second is the connection between the *personal* good and the *general* good.

Thus Broome considers how an additive theory might accommodate the value of *equality*. Specifically, he considers how an additively *separable* theory might accommodate equality. Thus he doesn't, contrary to Temkin, see *inequality* as an impersonally defective *pattern* in a distribution of individual wellbeing levels at a particular outcome.[83] Rather, he sees inequality as something that is bad in a way that relates directly *to the individual* whose wellbeing level is lower at a given outcome when someone else's wellbeing level at that same outcome is higher. Both of those negatives—*both* the lower wellbeing *and* the inequality—might then be registered, according to Broome, in the *personal good* that we calculate for that person at that outcome.

In other words, *if* we think that inequality is morally significant, we have the option of understanding the *personal good* as reflecting, not just the bare fact that a person has a *lower wellbeing* level at a given outcome, but also the fact that that person can be considered a victim of a *failure of equality* (or of *fairness* or of *priority*) at that outcome.[84]

Thus the concept of the personal good is *highly* adorned and *highly* accommodating.

The second step, then, is to make the connection between the *personal* good and the *general* good—that is, to use the concept of the personal good to transport the values of fairness, equality and priority into the additive picture. The value of equality is first captured in the concept of the *personal* good. It then exerts its

---

[83] Thus Temkin's hybrid, or pluralistic, theory might be *additive* in nature, but it's not *separately* additive: an impersonal pattern of inequality might detract from the value of a world, for Temkin, even if no person within that world can be counted in some sense a *victim* of that inequality.

[84] Fred Feldman has suggested a similar approach. Thus, in connection with the evaluation of outcomes, he proposes that the utility that an individual's existence contributes to the good of the outcome can itself be adjusted to take into account, e.g., justice. See *Pleasure and the Good Life*, pp. 195-197; and "Adjusting Utility for Justice: A Consequentialist Reply to the Objection from Justice," *Philosophy and Phenomenological Research* 55(3) (1995): 567-585.

influence on the evaluation of the outcome, or world, under scrutiny—the determination, that is, of the *general* good—via *summation*.

The upshot is what I will call P*.[85]

> P*. Where U(w) is the general utility of an outcome (or prospect) w and understood to represent the *general* betterness order between outcomes, where $u_1(w)$ . . .$u_n(w)$ are the personal utilities of the people in w, where a person's utilities "are defined to represent the person's betterness order" understanding betterness for the person as *personal* betterness, $U(w) = u_1(w) + u_2(w) + \ldots + u_n(w)$.[86]

To calculate the general utility of the world we simply add up the personal utilities that correspond to the personal good levels for each person who does or will exist in that world.

* * *

A theory can thus include the explicitly additive P* yet still recognize the value of equality since there's nothing in P* or any other component of Broome's framework that requires that the utilities to be summed up to determine the *general* good of a given outcome are the utilities that correspond to comparisons of people's *wellbeing* levels across a range of outcomes. Rather, the utilities to be summed up may be understood to correspond instead to a more complex comparison. Thus a person p's *wellbeing* may be the same in w1 as it is in w2 but w1 might still be *personally better for p* than w2, if, e.g., due to variations in the "conditions" of *other* people in w1 and w2, p has as much wellbeing as other people have in w1 but less than they have in w2.

11.2 *Narrow neutrality, the personal good and inversion.* It's Broome's concept of the *personal good* and the connection he makes between the personal

[85] Here I closely follow Broome's own description of Harsanyi's principle. See note [12] above [Harsanyi's].

[86] Broome 2015, pp. 250-251.

good and the general good that leads Broome to think that that P* does not rule out the values of equality, fairness or priority.

Does P* nonetheless rule out our *existential* values? Does it force us to say—on grounds entirely independent of the simple inconsistency argument and the assumption that w3 is surely not *worse* than w1—that w3 is *better* than w1? Does it rule out *narrow* neutrality?

Broome doesn't explicitly consider that question. I will just start by noting why we might think P* *doesn't* rule out our existential values.

Let's go back to the three outcome case. Per narrow neutrality, we say that Paula's existing at w3 does not make w3 *better* than w1 but that her existing at w2 does make w2 *worse* than w1.

The question now is whether Broome's additive framework rules out those happy results. Or, instead, can that same additive framework be understood to support those happy results? Can Broome's *highly* accommodating notion of the personal good be understood to reflect our *existential* values, just as, according to Broome, thinks it can be understood to reflect, e.g., our *egalitarian* values?

It seems, on the face of things, that it easily can. We simply take the position that Paula's *personal good* at w3 in point of fact falls at the single, sharp, neutral level despite the fact that her *wellbeing* level in w3 is positive. We can say, that is, that Paula's existence at the neutral level in w3 contributes *exactly* as much to w3's general good as Paula's never existing at all in w1 contributes to w1's general good—which is, of course, *none at all*. Summing up the relevant utilities—the utilities that correspond not to wellbeing but to the personal good—we then say that w3 and w1 are *equally generally good*.

To complete our account of the case, we take the position—indeed, must take the position—that, in addition, Paula's personal good at w2 falls in the *negative* range—again, despite the fact that her wellbeing level at w2 is itself positive. Summing the utilities now for w2, we say that w2 is *generally worse* than either w1 or w3.

In this way we can retain *both* narrow neutrality *and* P*. Of course, the account of the case we've just laid out commits us to a certain *inversion* in what

might otherwise have seemed a natural way of understanding the personal good. It means that personal good levels will fall either at the *none at all* level or at the *negative* level. If the former, bringing the additional person into existence *doesn't* make the outcome *better* even if wellbeing is positive. If the latter, bringing the additional person, whether at a *negative* wellbeing level (as in the miserable child case) or at an *avoidably* low positive wellbeing level (as in the case of Paula at w2), may well make the outcome worse.

But there's no reason to think that inversion is problematic. Indeed, to declare inversion out of bounds from the start—without, that is, argument—would in effect beg the question against narrow neutrality. Inversion is the mechanism that allows us, within the additive framework, to retain narrow neutrality. Inversion *is* narrow neutrality. Moreover, it's quite sensible. A *negative* personal good level *doesn't* mean that the person's life isn't worth living. It just means that the world itself is *defective* in some morally significant way that is rooted in how an existing or future person at that world fares. And surely such a world *is* morally defective. Consider w2. More can be done there for Paula than has been done at no cost to anyone else at all; w2 thus plausibly is the morally lesser world and the wrong choice.[87]

Thus it seems on the face of things that we can readily understand Broome's framework and specifically his concept of the personal good, as accommodating, not just the values of equality, fairness and, perhaps, priority, but our existential values—that is, narrow neutrality—as well.

11.3 *Two arguments against narrow neutrality*. Broome's overall framework—including P*—now before us, we are in a position to try to identify further arguments that would help Broome target not just the neutral *range* claim

[87] This Pareto principle needs to be spelled out very carefully. In the three outcome case, more is done for Paula at w3 than at w2, indicating a morally significant defect in w2. If we changed the case, and included a Quintus in w3 whose wellbeing in w3 is lower than it is in, say, some w4, then the condition on this simple Pareto principle would be failed: that would not be a case in which more can be done for Paula "at no cost to anyone else" since there would be a cost to Quintus in w3 notwithstanding the fact he never exists in w2.

but also *narrow* neutrality, and specifically, the claim that w3 and w1 are equally good. It might be tempting not to do that work and just to cheer narrow neutrality on. But we really can't comfortably *retain* narrow neutrality without first trying to identify and then evaluate Broome's arguments *against* narrow neutrality.

Broome's text suggests two such arguments. We'll start by briefly noting both arguments. We'll then examine each in more detail.

The first argument simply (i) looks at Paula's high wellbeing level in w3, (ii) notes that the position that her personal good level in w3 is *neutral* would mean that Paula's existence in w3 contributes just as much to the overall good of w3 as her never existing at all does and (iii) concludes that surely her personal good level in w3 must be at least a *little* greater than that!

Perhaps it's obvious that adding more detail to this first argument is not going to mean that it's not question-begging. We will consider what that detail might look like in what follows. But we may as well note now that the argument is likely *not* one Broome means to suggest.

The *second argument* may seem more promising. It starts with (i) the rejection of the neutral range claim. The argument thus takes as its first premise the conclusion of the simple inconsistency argument against the neutral range claim. So far so good. We then note that (ii) since the neutral level is, at most, a single, sharp, boundary, the odds are surely very much against anyone's ever coming into existence at exactly *that* level. Hence the odds are very much against Paula's coming into existence at exactly *that* level in w3. Put another way: the odds are very much against Paula's existing at a *wellbeing* level of +10 in w3 itself means that her *personal good* at w3 contributes nothing at all to the general good of w3. The odds are indeed so low as to justify our ignoring the possibility altogether. The upshot? Paula *doesn't* exist at the single neutral level in w3. Given, then, the relation between the personal good and the general good—given, that is, P*—we conclude that w3 and w1 therefore *aren't,* after all, equally good.

But that second argument seems to fail as well. What makes +10 neutral *is the case*. In particular, it's that +10 is *maximizing* for Paula within the context of *the particular case*, that is, the *three outcome case*. The narrow neutralist thus

would consider it no coincidence at all that the neutral level for that case and for that person would turn out to be +10. Anything less than that would be a *negative*; anything more than that would be *another case altogether*.

11.3.1 *First argument*. Let's take a closer look at the argument that I think Broome would *not* stand by. I did not, however, draw the argument out of thin air. There are some textual hints that favor it.

In the course of his discussion of the neutrality intuition, Broome underlines two closely related points. A person's existing at a *neutral level* at a given outcome means that that person's existing at that level contributes *exactly* as much to the general good of that outcome as that person's never existing at all at an outcome contributes to the general good of *that* outcome—that is, *none at all*. And, second, *any* level of the personal good—which is, just to underline, distinct from wellbeing; distinct, that is, from whatever it is that makes w3 better *for Paula* than w2 (or indeed w1) is—that is, in even the *slightest* degree, above the sharp boundary of the neutral level at a given outcome is such that the existence of a person at that level at that outcome will—by implication from P*—make that outcome generally better.

The text shows these points are central to Broome's discussion. Given the conclusion we are now after—that w3 and w1 are not equally good and that narrow neutrality is false—one might think the next natural step in the argument would then be just this: in view of Paula's high wellbeing level in w3, surely her existence at w3 *must* come with a personal good level that exceeds *by at least some slight degree* the sharp boundary of the neutral level. Hence, by P*, adding Paula to w3 after all makes w3 better than w1.

But if that's indeed the argument, then the argument fails. Why should the narrow neutralist accept the claim that Paula's existence at w3 exceeds by *any degree at all*, slight or not, the single, sharp boundary of the neutral level? Certainly, that claim can't serve as an *assumption* of the argument. As an assumption, it's obviously question-begging. After all, the very issue we are trying to settle is whether Paula's existence at w3 makes w3 better, which itself, under P*,

reduces to the issue of whether her personal good level at w3 exceeds the neutral level. Narrow neutrality claims that it doesn't.[88]

Perhaps, though, we can go still deeper and unearth a sub-argument for the otherwise question-begging claim that Paula's existence in w3 exceeds at least by a slight degree the neutral level. By hypothesis, Paula's *wellbeing* levels in w2 and in w3 are both positive. She is sufficiently well off in both w2 and w3 that the issue of whether her existence has exactly the same value *to her* as her never existing at all would have had is settled; her existence, in both w2 and w3, from her own point of view is *well* worth having. Surely, then, her wellbeing level at w2, though lower than it is at w3, cannot be *so* low that her existence in w2 makes w2 generally *worse* than w1 is. Surely, in other words, w2 is *at least as good as* w1 even if *not better*. That would in turn mean that Paula's personal good level at w2 *cannot* itself fall into the negative range—that is, that the personal good her existence in w2 contributes to the general good of w2 via P* cannot be *less* that the personal good her never existing at all at w1 contributes to the general good of w1. But if that's the situation with her personal good level at w2, then—since it might seem that we can surely agree that her personal good level in w3 exceeds her personal good level

---

[88] The argument, in other words, *assumes* that Paula's personal good in w3 itself is (at least very slightly) above the neutral level. But to *assume* or *stipulate* that Paula's personal good level in w3 is above the neutral level (or to claim it's positive on the grounds that her wellbeing level is positive or, indeed, let's suppose, at +10 *very* high) would be problematic (question begging) given the close (defined) relation between the personal good and the general good. After all, the question now on the table just is whether Paula's existence in w3 make w3 generally better than w1 is.

It is clear that Broome isn't aiming to foist off on us a question-begging argument favoring what seems to be his own clear conclusion on the neutrality intuition in general and extinction in particular: that is, that the neutrality intuition is false, and that it's not the case that extinction is neutral. At points, at least, he explicitly says that it's up to us to determine whether existence itself falls above or below or at the neutral level. ("Conceivably future people would on average life at the neutral level, in which case their existence together would be neutral. But that is such an unlikely coincidence we can ignore it. So the absence of all those future people will be either [personally and generally] good or bad. . . . I will leave this question unanswered." Broome, Toronto talk notes, p. 9.) Having been invited, then, to weigh in, narrow neutrality then does just that: despite her high wellbeing level in w3, Paula's personal good level in w3 is itself exactly neutral: w3 is neither generally better nor generally worse than w1 is.

in w2; the relation between wellbeing and the personal good may be complex but it's not *that* complex—we can infer that her personal good level at w3 after all *does* exceed the neutral level.

But that sub-argument fails. For—as noted earlier—to accept narrow neutrality is to accept *inversion*. It's just to accept that Paula's level of the personal good in w2 falls *below* the neutral level—falls, that is, into the negative range—*despite* the fact that her wellbeing level at w2 is unambiguously positive. Thus we may well agree that Paula's personal good level in w3 exceeds her personal good level in w2. But it's not going to follow that her personal good level in w3 exceeds the neutral level—that, in other words, her existence in w3 makes w3 better than w1 is.

Left without any adequate sub-argument for the claim that Paula's existence in w3 exceeds the neutral level, we should reject the first argument as question-begging. As noted before, it's doubtful that that first argument is one Broome meant to put forward to begin with.

11.3.2 *Second argument*. Let's now take a look at the second argument—the one I think Broome may well stand by.

The second argument begins with the conclusion of the simple inconsistency argument against the neutral range claim. The second argument thus starts with the point that the neutral level constitutes at most a single, sharp boundary—that is, that it can't be the case that *both* w2 and w1 are equally good and that w3 *and* w1 are equally good. We then simply note that any given person might exist at any one of perhaps infinitely many possible wellbeing levels. Paula's existing at +5 in w2 and at +10 in w3 are just two of those many levels. We then consider the odds against the proposition that her existence at +10 in w3 happens to coincide with *exactly that* single, sharp level. Surely they are very small—so small that we can safely "ignore" them altogether.[89] That in turn would mean that

[89] "Conceivably future people would on average life at the neutral level, in which case their existence together would be neutral. But that is such an unlikely coincidence *we can ignore it*. So the absence of all those future people will be either [generally] good or [generally] bad." Broome, Toronto talk notes, p. 9 (emphasis added).

Paula's existence at w3 "will be either [generally] good or [generally] bad."[90] We've already accepted as an assumption that Paula's existence in w3 doesn't make w3 *worse* than w1 (we are not Benatarians).[91] We thus conclude that Paula's existence at w3 makes w3 generally *better* than w1 is.

Does this argument work? Can we on the basis of statistics dismiss the possibility that Paula's existence at w3 is itself neutral?

Let's step back. It seems clear that Broome's statistical argument does not even begin to look viable *unless* we eliminate the restriction that I included in my original presentation of the three outcome case and my original presentation of Broome's argument—that is, the *intra-case* restriction. According to that restriction, the three outcomes displayed in the three outcome case exist as *accessible* outcomes, and we are asked to compare w3 against w2 on the assumption that w3 isn't simply a remote *logically possible* world but rather an *accessible* world, a possible future agents had the ability, the power, the resources to make happen.

Moreover, the Pareto-like principle, articulated to include that same restriction, strikes us as simple, straightforward and indeed compelling. When w3 and w2 exist as *accessible* outcomes, we all agree that w3 is *morally better* than w2.

On the basis of that work, we then easily agreed that the neutral range claim is inconsistent.

We then pointed out that we could avoid the inconsistency by taking the position that w2 is worse than w1—and that we would then be free also to say that w3 is equally as good as w1. In other words: *within the particular case and for the particular person*, Paula, there exists at most a single, sharp neutral level of existence, that is, +10 at w3.

---

[90] Broome, Toronto talk notes, p. 9.

[91] See note 28 above.

In fact, however, there is strong textual support[92] for the notion that the three outcomes Broome describes for purposes of constructing his inconsistency argument against the neutral range claim are *not* meant to be assumed to exist as accessible outcomes within the context of a given case—that is, that that argument is meant to proceed *without* any reliance on the intra-case restriction. The conclusion of the argument would then be considerably *stronger* conclusion—that, *for all outcomes whether accessible or not and (perhaps) for all people*, there exists at most a single, sharp neutral level of existence.

Adopting the intra-case restriction, we understand the argument to unfold within the confines of a particular hypothetical—a single case in which the issue is whether a given person is to be brought into existence at one of two distinct wellbeing levels or not brought into existence at all. Presented with that hypothetical, we are willing immediately to agree that w3 is better than w2 is.

But then—as noted above—we're not going to consider against all odds that Paula's existence at +10 in w3 would happen to fall at the single, neutral level. What makes +10 neutral *is the case*. Paula's existence at +10 in w3 represents the *best* that can be done for Paula *in that case*; w3 is *maximizing* for Paula *in that case*. So of course *for that case* the neutral level will turn out to be +10, exactly the level at which Paula exists in w3.[93]

But now we are considering an alternate construction of Broome's argument. We are now considering the possibility that Broome means for us to *drop* the intra-case restriction—that he means his argument to reach for the

---

[92] Broome himself states at one point that that is exactly the conclusion he means to reach. "But when we evaluate B in comparison to C, we must not assume B and C are actually available alternatives. Nothing says they are." Broome 2004, p. 147.

[93] A point of clarification. It's true that *in one sense* there exists, for all people and all cases, a single neutral level: that is, that for all people and all cases, by definition, a person's existence at the neutral level in an outcome *by definition* contributes exactly as much personal good to that outcome as that the personal good a person's never existing at all in an outcome contributes to that outcome. The *neutral* level of the personal good is just the *none at all* level. But to think that that point itself means that there exists a single neutral level of *wellbeing* would be to confuse the *personal good* on the one hand and *wellbeing* on the other. That's a confusion that practically calls out to be made but it's a confusion all the same.

conclusion that, for all people and all cases, there is a single, sharp, neutral level of existence.

That, in turn, would mean that, as we proceed to compare w3 against w2, for all we know, w3 hails from one case—the *three* outcome case—and w2 from different case altogether, a case in which w3 *does not exist* as an accessible outcome, that is, the *two*-outcome case.

Graph 11.3.2: Two Outcome Case	Wellbeing	w1	w2
	+10		
	+5		**Paula**
	+0	*Paula**	

We are then asked to accept the claim that w3 in the three outcome case is better than w2 in the two outcome case.

Now, on this *inter*-case construction of Broome's argument, the point about just what a wild coincidence it would be for Paula's existence at +10 in w3 to fall at the single, sharp neutral level comes into play. For there will always be still another case—a case involving, e.g., an outcome w4 in which Paula's wellbeing level is greater than it is in w3, and a case involving outcomes w4 and w5 in which Paula's wellbeing level is greater in w5 than it is in w4 and greater in w4 than it is in w3. And so on. Given that potentially endless array of levels of wellbeing at which Paula *might* come into existence, why should we think that there's any real chance at all that Paula's existence at +10 in w3 would happen to fall at the single, sharp, neutral level?

Dropping the intra-case restriction thus may make Broome's statistical argument begin to look potentially viable.

In fact, however, there are still difficulties. We can reject the intra-case restriction and accept that the chances are very much against Paula's existence at +10 in w3 falling just at the neutral level. But the moment we reject the intra-case restriction the argument becomes vulnerable at another point. *Intra*-case restriction

in place, the argument against the neutral range claim can proceed on the basis of a Pareto-like principle that is, due to the restriction itself, simple, straightforward and compelling. When we understand that the case includes as accessible outcomes both w2 and w3, we are happy to say that w3 is morally better than w2 is. Without that restriction in place, the Pareto-like principle the argument asks us to put to work becomes much stronger. We are asked immediately to draw the inference that, even if w3 *doesn't* exist as an accessible alternative to w2, w3 is still *better*—*morally* better—than w2 is.

Now, on the face of things, that result may not seem objectionable. In fact, however, it's a result that is anathema to narrow neutrality. For given narrow neutrality's prior commitment to the position that w1 and w3 are equally good, this new result would then commit narrow neutrality to the position that w2 is *worse* than w1 in the *two* outcome case. But that's just not a plausible position. (Again, we are not Benatarians.) The reason adding Paula to w3 in the three outcome case doesn't, according to narrow neutrality, make w3 worse, or better, than w1 is just that Paula's wellbeing level in w3 has itself been maximized. That same reasoning applies to w2 in the two outcome case.

We shall thus want to say—though in a more exacting vocabulary[94]—that w2 being worse than w1 in the three outcome case does not imply that w2 is worse than w1 in the two outcome case. We shall, in other words, want to reject the stronger, inter-case version of the Pareto-like principle.

This point can be made entirely without reference to whether our Pareto-like principle is to be understood to be limited to the case where the outcomes we are ranking, w2 and w3, hail from the same case or from two different cases altogether. For purposes of developing the inconsistency argument against the neutral range claim, we are willing, whether on Pareto-like grounds or on other grounds entirely, to accept that w3 is better than w2. (Our maximizing intuitions are at play when we do that—but we can certainly get to that result without thinking that what is to be maximized is wellbeing in the *aggregate*.) But now let's make it explicit that the w2 we are asked to compare against w3 doesn't have w3 as an

[94] See part ___ below.

accessible outcome. There is, in other words, no w3 to exert a deflationary effect on w2; we have no grounds for saying w2 isn't equally as good as w1. There is no w3 such that we can say that w3 shows that w2 is worse than w1 is. On those facts, we are no longer willing to accept that w3 is better than w2 is.

There is still another problem with Broome's argument. We earlier conceded that surely w3 wasn't worse than w1. But if we accept the unrestricted Pareto-like principle, that concession shall need to be clawed back. If how w3 compares against w1 is to be determined, not by reference to the outcomes that exist as accessible to w3, but by reference to *all possible outcomes*, then given that, for any particular wellbeing level Paula has in any particular world, there is some possible world such that Paula's wellbeing level is at least a little higher, narrow neutrality would imply that all such worlds are actually *worse* than Paula's never existing at all. Broome might argue that that result shows that narrow neutrality cannot itself be correct. But we can't validly infer that result. For it's just as plausible—and indeed the position of the narrow neutrality—that how w3 compares against w1 is to be determined, not by reference to all *possible* outcomes, but rather by reference to all *accessible* outcomes—all outcomes, that is, that exist as accessible outcomes within the context of the particular case.

11.4 *Looking ahead*. What we are in effect saying here is that, in the three outcome case, w3 is better than w1 but that, in the two outcome case, w2 and w1 are equally good. But that way of looking at our facts may itself seem highly objectionable. It may seem to *over*-contextualize the discussion; it may seem to force us to the result that whether w2 is just as good as, or worse than, w1 can *vary* depending on the case, a result that in turn raises a host of theoretical issues. Thus we will need to make our substantive point in a considerably more exacting vocabulary. In the next chapter, we will thus attend to that issue and along with a handful of others. Assuming those issues can successfully be addressed, however, we will then be in a position to conclude that Broome provides us with a compelling argument against the neutral *range* claim but no effective argument at all against *narrow* neutrality.

Chapter 12

Objections and Replies

12.1 *First objection*. If Paula's personal good in w2 is negative in the *three* outcome case, it must be negative in w2 in the *two* outcome case as well. That would mean, in turn, that it, after all, makes things *generally worse* to bring Paula into a perfectly fine existence in the two outcome case. However, while it's plausible to think that w2 is generally worse than w1 in the *three* outcome case, it's not at all plausible that w2 is generally worse than w1 in the *two* outcome case.

Reply: Narrow neutrality rejects the claim that, if Paula's personal good in w2 is negative in the *three* outcome case, it's negative in w2 in the *two* outcome case. Her *wellbeing* stays *constant* from one outcome to the other but consistent with that point her *personal good* in w2 in the two outcome case may be *greater* than it is in w2 in the three outcome case. That is, her personal good in w2 in the *two* outcome case may be exactly the same as her *personal good* level in w3 in the *three* outcome case (her wellbeing level having been maximized both in w2 in the two outcome case and in w3 in the three outcome case; there being nothing in the two outcome case to exert the same deflationary effect on her personal good level in w2 that w3 exerts in the three outcome case).

In taking the position that Paula's personal good in w2 in the two outcome case is the same as her personal good in w3 in the three outcome case—and that, correspondingly, the general good of w2 in the two outcome case is the same as the general good of w3 in the three outcome case—we remain in compliance with the rule that "[t]he value of a distribution depends only on the condition of each person; that is a consequence of the principle of personal good . . . . If the presence or absence of alternatives affects the value of a distribution, it can do so only by affecting some person's condition."[95] Thus I am not proposing that we calculate the personal good and then, depending on what outcomes are accessible, *reduce* the value of w2 in the three outcome case. Rather, I am proposing that in the three outcome case we build the deflationary effect of w3 into Paula's personal good at w2 . . . and *then* determine the value of the outcome w2. In the two outcome case,

[95] Broome 2004, p. 147.

there is no such deflationary effect on w2. So there is no basis on which to say w2 in that case is generally worse than w3 in the three outcome case.

We are here, in effect, *contextualizing* the value of Paula's existence in w2 and—with that, given P*—the evaluation of w2 itself. That is: the claim is that we can't fully assess the value of Paula's existence in w2, or determine whether w2 is worse than w1 or equally as good as w1, until we know what case we are in—until, that is, we know whether or not w3 exists as a further accessible outcome.

Broome himself notes that we can say that Paula is "wronged"—that is, her wellbeing level is avoidably reduced in a case where increasing her wellbeing could have been achieved at no cost to anyone else—in w2 in the three outcome case. That fact may provide us with grounds, as in the case of an inequality, for "reducing the value" of w2.[96] But we have no such grounds in the two outcome case. Hence we shouldn't, in that latter case, consider the value—that is, the general good—of w2 "reduced."

11.2 *Second objection*. w2 in the two outcome case *can't* be generally better than w2 in the three outcome case. Nothing relating to Paula's existence or the existence of anyone else has changed from one outcome to the other.

Reply: It's true w2 in the two cases is the same *for Paula from Paula's own point of view*—that is, that Paula's *wellbeing* level in w2 is the same in the three outcome case as it is in the two outcome case. But just as *wellbeing* can be the same for a subject in a given outcome from one case to another but due, e.g., to an unfairness, the outcome in the one case can be *personally better* for the subject than in the other case, *wellbeing* can be the same for Paula in w2 in both cases but w2 in the two outcome case can still be *personally better* for Paula than w2 in the three outcome case.

Now, this reply itself opens to the door to still another objection, an inconsistency objection. If Paula has a certain amount of personal good in w2, how can have more personal good than that in w2? How can she have more personal

[96] Broome 2004, p. 147.

good in w2 in the two outcome case than she has in w2 in the three outcome case? We'll set this objection aside for now and return to it in part 12. 4 below.

11.3 *Third objection*. The dire facts I have built into the three outcome case are at odds with the ordinary case Broome has in mind when he refutes the neutral range claim. As I have constructed that case, Paula's personal good in w3 is "none at all" and in w2 it's actually negative. Broome, though, would have mentioned it if he had meant the case to include such dire facts!

Reply: Broome can't mean for us to *stipulate* as part of an argument against the neutral range claim that Paula's *personal good* in w2 and in w3, or even just in w3, is positive. Given the relationship between the *personal* good and the *general* good, such a stipulation would be problematic. After all, our question just is: does adding Paula at w3 (or w2) make that outcome generally better? Making it part of the original set up of the case that Paula's personal good in w3 (or w2) is *positive* would blatantly beg the question.

Moreover, once we distinguish wellbeing and the personal good, there's no basis for describing the facts I have built into the case as at all "dire." If characterizing her *personal good* as none at all at w3 and actually negative at w2 may seem a little glass-half-empty-ish—or as Johann Frick has put the point *harsh*—one can feel free to use the term *contributory value* in place of *personal good*.

11.4 *Fourth objection*. Narrow neutrality proposes what may seem to be a violation of the axiom of the independence of irrelevant alternatives, that is, the *independence axiom*. How can the mere *accessibility* of w3 leave w2 worse than w1 in the three outcome case but equally as good as w1 in the two outcome case?

In this connection, we also face a *consistency* question, one we deferred above: if Paula has a certain amount of personal good in w2, how can she at the same time have *more personal good than that* in w2? How can she have more personal good in w2 in the two outcome case than she has in w2 in the three outcome case?

Reply: Our discussion here can be brief. For this particular inconsistency argument is one that we have already seen. Thus, in the context of our discussion in Part I of whether any correct solution to the nonidentity problem was bound to abide by a certain axiological constraint, we considered whether the view that a comparison of one world w1 against a second world w2 in some cases depends on facts relating to still a third world w3 is consistent. There, as here, any such w3 that might affect, indeed, change how w1 compares against w2 will itself be an *accessible* world. But w3's accessibility relative to—say—w2 is a feature that can itself be discerned upon careful examination of w2. There is, we said, going to be a causal explanation of w3's accessibility—an explanation that is itself rooted in the *modal details* inherent in w2: how things, within the bounds established by, e.g., the laws of nature and (perhaps) the acts of other agents, *could* have been. By the same token, when w3 isn't accessible relative to w1, that, too, is going to have a causal explanation, itself rooted facts about w2. To say that w3 isn't accessible in that case is just to say that agents in w2 *lacked* some power, some ability, to make things any better for p, that is, to bring about an alternate possible future that includes the advantages for Paula we see in w3.

Having come this far, we can then easily see our way clear to the next step. That agents have the relevant ability in the one world and lack that ability in the other just means that those worlds—w2 in the three outcome case and w2 in the two outcome case—are actually two *distinct* worlds. Worlds, after all, aren't simply *distributions*—bare boned assignments of wellbeing levels to members of a particular population. Rather, worlds come to us with all their details necessarily intact. New details entail new worlds.

A more exacting vocabulary will recognize exactly that point—and the inconsistency we were worried about never in fact arrives. Thus we might say that Paula indeed has less personal good in w2 than she has in w1 in the three outcome case due to the accessibility in that case of w3, which accessibility is itself reflected in the modal details inherent in w2 and w1. But she has exactly the same amount of personal good in w2′ than she has in w1′ in the two outcome case—and this last,

despite the fact that w2′ and w2 distribute wellbeing across exactly the same population in exactly the same way.

What secures this result—guaranties, that is, that we won't come across still another case a case in which we seem bound to recognize an identity between w2 and w2′ and hence face inconsistency all over again—is the *accessibility axiom*.

> *Accessibility axiom*. If wβ is accessible to wα, then *necessarily* wβ is accessible to wα.

Avoiding inconsistency by introducing a more exacting vocabulary means that we can retain the independence axiom understood in a certain way. If the principle means we aren't allowed to look closely enough *at w1 and w2* to see whether *w3* is indeed accessible—if it means we must blind ourselves to those particular facts about w1 and w2—then independence must go. But if it's understood as imposing not *quite* such a ridiculously strict standard as that, then it may stay.

Chapter 13

Conclusions; Implications

13.1 *Inversion and narrow neutrality.* A main purpose here has been to suggest that we might save the only version of the neutrality intuition that we want to save—that is, *narrow neutrality*—through an *inversion* of the picture we perhaps first imagined when presented with the neutrality intuition. Rather than thinking of differences in wellbeing levels as having some effect on general good in some range *above* the neutral level, we can think of those differences as having some effect on the general good in some range *below* the neutral level.

The tension all along for the neutralist—that is, for the theorist who wants via some form of the neutrality intuition to take existential values into account; that is, for the *existentialist*—is to explain why Paula's higher wellbeing level in w3 has a critical impact when we compare w3 against w2 but no impact at all when we compare w3 against w1. Why is it *something* when we compare w3 against w2 and *nothing* when we compare w3 against w1? Why is her higher wellbeing level so potent when it comes to the one comparison but so completely ineffectual when it comes to the other? How can a mere shift in the *question we happen to be asking* change the *value* of Paula's higher wellbeing level? How can Paula's higher wellbeing level in w3 make w3 better than w2 *without* making w3 better than w1?

The *inversion* that the neutrality intuition puts into place resolves that tension. It does so by setting the personal good level for Paula's existence at w3 at the *none at all* level—at, that is, the level that is exactly the same as the level of personal good Paula's never existing at w1 contributes to the general good of w1—and setting the personal good level for Paula's existence at w2 *below* the neutral level. That means that the personal good Paula has at w3 can stay perfectly constant at the *none at all* level whether we happen to be comparing w3 against w2 or happen to be comparing w3 against w1. And that in turn means we can say—in fact it *requires* us to say given P*—that her existing at just that level in w3 indeed makes w3 better than w2 but does not make w3 better than w1.

A related tension arises in connection with our evaluation of w2. Why is Paula's existence in w2 perfectly innocuous when we compare w2 against w1 but

morally troubling when we compare w2 against w3? Larry Temkin attempts to resolve that tension by bringing into the analysis of the three outcome case two very distinct sorts of views, the *Internal Aspects View* and the *Essentially Comparative View*. The former suggests that w2 is at least as good as w1 is (based on the internal aspects of each of the two outcomes and without reference to what is going on in any third outcome) while the latter suggests that w2 *isn't* at least as good as w1 is (based on an examination of outcomes *beyond* w1 and w2; based, that is, on the accessibility of w3).

I see narrow neutrality as supportive of what Temkin is aiming to accomplish in taking the position that the evaluation of the three outcome case implicates both the Internal Aspects View and the Essential Comparative View. In contrast, though, to Temkin's approach, which ultimately involves weighing a plurality of values against each other, narrow neutrality, by setting (in the three outcome case) Paula's existence in w2 below the single neutral level whether we are attending to w3 as an available alternative outcome or not, will exclude the problem result: that is, the result that w2 is at least as good as w1 is. (Temkin himself, I should note, may view that point as a deficiency in, not an advantage of, narrow neutrality.)

13.2 *The nonidentity problem*. In Part I, I argued that the nonidentity problem did not present, on further examination, a problem for the person-affecting intuition, that is, the intuition that a "bad" act must be "bad for" some person or another who does or will exist. There was a proviso on that conclusion. For it to hold, the person-affecting intuition had to be understood in a certain way. PAIA(c) wouldn't do; we needed PAIA* instead.

It's interesting now to see how my proposed solution to the nonidentity problem, which insisted that we take into account, in determining whether a given act is "bad for" a given person or not, the full modal array, and not some arbitrarily limited subset, can be expressed under a theory of narrow neutrality.

Thus we can say that Andy and Rachel exist at the neutral level in w2 and w3, respectively, but that Andy exists below the neutral level in w1. Thus the

personal good contributed by Andy's and Rachel's existence in w2 and w3 to the general good of w2 and w3 is at the "none at all" level, whereas Andy's personal good in w1 actually falls into the negative range, meaning that his existence there actually takes away from the overall good of w3.

Thus, where (here and in what follows) *pg* stands for *personal good*, and n is a positive number:

Graph 13.2: Narrow Neutrality and the Modally Enriched Presentation of the Nonidentity Problem

wellbeing	w1 (including a1)	w2 (including a2)	w3 (including a3)
+10		**Ruth** pg = 0	**Andy** pg = 0
+8	**Andy** pg = -n		
+0	*Ruth**	*Andy**	*Ruth**

13.3 *Infinite population problem*. If we overlook the possibility of inversion we limit our capacity for cogent analysis. By restricting our answers to how much Paula's situation adds to the overall good of each outcome to the *positive*—if we think of it as positive in w3, positive, though lower, in w2 and *none at all* for w1—we handcuff ourselves. If we instead recognize the contribution the addition of Paula makes to w2 as *below* the neutral level, we become able to say sensible things about that case.

The inversion we obtain from narrow neutrality doesn't just help us in the three outcome case. There are other cases as well in which that strategy facilitates analysis. Consider, e.g., the *infinite population problem*. In that problem, we are to imagine an infinite population existing at a relatively low, though still clearly

positive, wellbeing level in w1, and that same population existing at a significantly higher wellbeing level in w2. Totalism, of course, immediately seems defeated by this case, since it counts w1 and w2 as equally good, whereas it seems intuitively clear that w2 is morally better than w1. That's especially so, if we do imagine how things look from the perspective of the individual members of the population.

More generally, the infinite population problem is difficult if we think of each of the infinite lives worth living as adding something *positive* to the overall good of the relevant outcome. We are tempted to reconstruct the case in a way that has us apply the additive principle not to the infinite set but rather to selected finite subsets of that infinite set. But if, for the outcome w1 in which each person's wellbeing level is, say, +1, we think of those person's lives as adding something *negative* to w1—if, that is, we think of the personal good level as falling into the *negative* range—and, for the outcome w2 in which each person's wellbeing level is +2, we think of those person's lives as adding *nothing at all* to w2, we then have a basis for the claim that w2 is generally better, indeed *infinitely* better, than w1 is—which is, of course, exactly what we would like to say about that case.

Again using grays to represent wellbeing at the neutral range, and reds to represent wellbeing that falls below the neutral range, we can sum up as follows:

Graph 13.3: Infinite Population Problem		
Wellbeing	w1	w2
+2		**p1, p2, p3….** (pg + 0)
+1	**p1, p2, p3….** (pg = -n)	
+0		

13.4 *Replaceability*. What about the poor little dog we thrust into the limelight in the introduction to this book and have since lost track of completely? Let Dolly be the little dog. The question is to compare an outcome in which Dolly's life is saved how the outcome in which Dolly's life is saved compares against the outcome in which Dolly is painlessly euthanized and a distinct little dog Jolly is created in the lab that is so similar to the original that—like a new pet minnow—the family itself will never know the difference.

I am taking for granted here that we reject the totalist account of this case. What we want to know, rather, is whether we can make sense of our intuition that w2 is actually better than w3 notwithstanding our commitment both to maximization and to impartiality. My view is that we can. While Dolly's wellbeing in w2 is identical to Jolly's in w3, it's also the case that Jolly's existence at w3 can plausibly be viewed, under narrow neutrality, as contributing nothing at all to the general good of w3—that is, that her personal good at w3 is +0. At the same time, under inversion, we shall want to recognize that Dolly's existence at w2 itself detracts from the general good—that her personal good level at w2 actually falls into the negative range. No one else being affected either way, we may immediately conclude that w2 is generally better than w3 is. What of w1? We are free to say there that Dolly's personal good level is, as it is in w3, in the negative range, and hence, again simply adding things up, that w1 is actually worse than either w2 or w3. Accepting the very close connection between betterness and what we ought to do, we conclude, finally, that the act that results in w2 is obligatory and the other two just wrong.

Summing up:

Graph 13.4: Replaceability (m and n are positive numbers and n > m)			
Wellbeing	w1 (we do nothing)	w2 (we cure Dolly)	w3 (we painlessly euthanize Dolly and create Jolly)
+8		**Dolly** (pg = +0)	**Jolly** (pg = +0)
+4			**Dolly** (pg = -m)
+0	**Dolly** *Jolly** (pg = -n)	*Jolly**	

Accounts of other problem cases in population ethics are explored in Appendix A below.

APPENDIX A

Key: Bold face means the indicated person does or will exist in the indicated outcome; italics/* means indicated person never exists in indicated outcome. Throughout, m and n are positive numbers and n > m.

Implications of *narrow neutrality* (in combination with other plausible principles) for levels of the personal good (*pg*) are shown in the graph.

Double Wrongful Life (Parfit)		
Wellbeing	w1	w2
+10	**George** (pg = +0)	**Jill** (pg = +0)
+9 … +2		
+1		**George** (pg = -n)
+0	*Jill**	

Adding up levels of the personal good to determine the overall good, we find that w2 is worse than w1.

Addition Plus			
Wellbeing	w1	w2	w3
+11		**p** (pg = +0)	
+10	**p** (pg = +0)		
+5			**p** (pg = -m) **q** (pg = +0)
+1		**q** (pg = -n)	
+0	*q**		

Assumption: In Addition Plus, priority view allows us to say that that p's existing in w3 (at wellbeing +5) takes less away from the overall good than q's existing in w2 (at wellbeing +1) does.

Mixed Existence (Tradeoff to Exist)			
Wellbeing	w1	w2	w3
+4		**q** (pg = +0)	**p** (pg = +0)
+3		**p** (pg = +0)	**q** (pg = +0)
+0	*p**, *q**		

Puzzle: In Mixed Existence, we want to say each of the three worlds is exactly as good as each other. We do not want the fact that p in w2 has less wellbeing than p might have had to compel us to conclude that w2 is worse than w1. So can't assign a personal good level below the neutral level to p in w2 (or to q in w3). The puzzle then is why it's consistent for us then to assign a personal good level below the neutral level to q in w2 in Addition Plus. The basis for that "negative" assignment in Addition Plus is that q exists in w2 and has less wellbeing in w2 than q has in w3. But that basis exists in Mixed Existence as well. Solution: there is a difference in the two cases. In Mixed Existence, the only way to make p better off than p is in w2 is to make q exactly as badly off as p is in w2. In Addition Plus, there is a way of making q better off than q is in w2 that does not make anyone as badly off as q is in w2.

Double Wrongful Life (Parfit)		
Wellbeing	w1	w2
+0	*p**	*q**
-10	**q** (pg = +0)	**p** (pg = +0)

Here, w1 is exactly as good as w2, which seems correct.

Multiple Wrongful Life			
Wellbeing	w1	w2	w3
+0	*p**, *q**	*p**	
-10	**r** (pg = -n)	**q r** (for each of two, pg = -n)	**p, q, r** (for each of three, pg = -n)

Note on multiple wrongful life: Retention of P* over the personal good means that if we multiply wrongful lives we will be making each successive outcome worse than the prior outcome.

Single Wrongful Life		
Wellbeing	w1	w2
+0	*p**	
-10		**p** (pg = -n)

Personal good of p in w2 in Single Wrongful Life is distinct from personal good of p in w2 in Double Wrongful Life even though wellbeing levels in the two outcomes in the two cases are identical. Why? Because personal good is determined by factors beyond simple wellbeing, including conditions of other people's existence (e.g. conditions of q's existence in w1 in Double Wrongful Life).

Tom, Dick and Harry (Parfit)			
Wellbeing	w1	w2	w3
+4	**Harry** (pg = +0)	**Dick** (pg = +0)	**Tom** (pg = +0)
+2	**Dick** (pg = +0)	**Tom** (pg = +0)	**Harry** (pg = +0)
+0	*Tom**	*Harry**	*Dick**

For same reason we conclude in Mixed Existence that we wouldn't assign personal good below the neutral level to p's existing in w2 or q's existing in w3, we want to say here as well that we won't assign personal good below the neutral level to Tom's existing in w2 (Harry's existing in w3, etc.).

APPENDIX B

First Argument: Broome's Argument Against the Neutral Range Claim

Let "=" between outcomes mean that each outcome *is equally as generally good as* the other; let "<" between outcomes mean that the first outcome *is generally worse than* the second; and let "<" otherwise have its usual meaning.

Consider a case in which a person Paula never exists in w1; exists and has m level of wellbeing in w2; and exists and has n level of wellbeing in w3. Assume that m<n and that w1, w2 and w3 are otherwise the same (same population; same distribution). Assume that w3 exists as an accessible outcome relative to w1 and w2. The question is whether Paula's existing at wellbeing level m in w2 and Paula's existing at wellbeing level n in w3 can *both* count as Paula's existing at the neutral level, i.e., as instances in which her existence does not make the outcome itself better (or worse).

1. In this single case, there exist at least two neutral levels m and n.	Assumption of neutral range claim for *reductio*; facts of case
2. w1 = w2.	(1), definition "neutral level"
3. w1 = w3.	(1), definition "neutral level"
4. w2 = w3.	(2), (3), transitivity and symmetry of =
5. w2 < w3.	Pareto-like principle (restricted to the case where w3 is accessible to w1 and w3)
6. Inconsistency	(4), (5)
7. The neutral range claim is false; m and n can't both be neutral levels.	Reductio (1)-(6)

Lines (1)-(4) seem unobjectionable. But a note on line (5) is in order. Given that m and n designate levels of wellbeing, not the personal good, and given the intra-case restriction—that we are in a case in which w3 exists as an accessible outcome relative to w1 and w2—(5) seems unproblematic. The upshot is that the argument goes through—that the neutral range claim (intra-case restriction in place) is false.

Broome justifies line (5) by reference to the principle of personal good (PPG). Despite its title, there really is a question whether PPG is intended to talk about simple *wellbeing* levels or levels of the *personal good*. Let's consider the first reading first. That reading makes PPG equivalent to the Pareto-like principle I have cited as justifying line (5).

> *The principle of personal good* (*PPG*). Where two outcomes have the same population, if one outcome assigns at least as much wellbeing to each member of the population as the other is and more wellbeing to at least

> some member of the population than the other does, then the one outcome is generally better than the other (WL p. 120, rewriting using GPG vocabulary).

Given PPG, so interpreted *and* taken together with the restriction I have included in the justification column for (5), line (5) seems unobjectionable. After all, PPG is *both* a "same-population" *and* a "same-person" principle. It's explicitly limited to the case where two outcomes share exactly the same population. And, again explicitly, the sufficient condition is satisfied only if the first outcome is personally better for a person than the second outcome is *for that same person*.

Given the inconsistency in (6), Broome rejects the neutral *range* claim in favor of the *single neutral value claim*:

> *Single neutral value claim*: There exists at most one neutral level of the personal good—one level of the personal good such that adding a person at that level makes an outcome neither generally better nor generally worse.

That is: for *any one person* within *any one case* that includes the details we have included here—same population; better for at least one and worse for none; no outcomes beyond the three outcomes described—there exists at most a single neutral level. With those caveats in mind, I am happy to accept the argument and the conclusion.

Let's now consider the second reading. To generate the second reading, we simply substitute personal good in for wellbeing throughout PPG.

But now we have an even quicker argument. If m and n talk about the personal good, and $m < n$, then just in virtue of how P* defines the terms it follows that $w2 < w3$. It can't be, then, that both m and n are neutral levels. So again the neutral range claim is defeated. But notably the account would leave two options open: either m falls into the negative range, or n falls into the positive range. Narrow neutrality would favor the former option. In any case, assuming the second of these two options holds would be to beg the question against narrow neutrality.

Second Argument: Argument Against Narrow Neutrality

1. There is only one neutral level.	Single neutral value claim
2. Adding Paula at the neutral level adds exactly as much personal good to an outcome as Paula's never existing at all adds to an outcome.	Definition "neutral level"
3. Adding Paula at any level of the personal good *greater* than the level of personal good that Paula's never existing at all adds makes an outcome generally *better*.	Lines 1 and 2 and P* (more personal good entails more general good)
4. Paula's personal good at level n is greater than the level of personal good that Paula's never existing at all adds to an outcome.	Facts of case (Paula's wellbeing level in w3 is +10; she has a very good life; not even close to the sharp boundary below which lives aren't worth living)
5. Adding Paula to w3 makes w3 generally better.	Line 4, relation between personal good and general good, P*
6. Personal good levels "often"—within limits—are greater than the level of personal good that a person's never existing at all adds to a given outcome.	Line 5, universal generalization (nothing special about this case)
7. The neutrality intuition is false; "often"—within limits—adding a person makes the world generally better.	Line 6

This argument I take not to be Broome's argument (though he might, recognizing (4) as a mere assumption, accept the *conditional* that, if (4) holds, adding Paula to w3 in the three outcome case makes w3 generally better). As the argument stands, however, we may reject the conclusion on the grounds that we have no basis on which to accept (4) and any stipulation that (4) holds would be question-begging against the view that adding Paula to w3 doesn't make w3 generally better.

APPENDIX C: *Procreative Asymmetry*

The *procreative asymmetry* consists in two highly plausible claims that together seem to lead us into inconsistency. According to one half of the asymmetry, agents, other things equal, are *not* morally obligated *to bring into existence* a well-off (for short, a *happy*) child. But, according to the other half of the asymmetry, agents, other things equal, *are* morally obligated *not* to bring into existence a child whose life is *less* than worth living (for short, a *miserable* child). Thus:

Procreative Asymmetry (n is a positive number)				
Wellbeing	a1 in w1: cause Meg to exist in w1	a2 in w2: cause Meg not to exist in w2	a3 in w3: cause Hans not to exist in w3	a4 in w4: cause Hans to exist in w4
+100				**Hans** (pg = 0)
0		*Meg**	*Hans**	
-100	**Meg** (pg = -n)			

> The Procreative Asymmetry consists in the following claims: a1 is wrong and a3 is permissible (on the deontic side) and w1 is worse than w2 but w3 isn't worse than w4 (on the telic side).

Whence the inconsistency? If it's morally important for agents not to make a child miserable, then it's morally important for agents not to *not* make a child happy. After all, as Singer and other committed consequentialists effectively argue, making a child miserable and *not* making a child happy

are just two ways of accomplishing (whether by act or omission) the same morally suspect end: making a child *worse off* than that child might have been. Moreover, if the *miserable* child, despite the fact that the existence of that child remains contingent (hinges, that is, on how the choice is made), has full moral status—if agents, that is, have obligations to make things better for rather than worse in respect of the child despite the fact that that child's coming into existence is exactly the choice under scrutiny—so, surely, does the *happy* child have full moral status. Hence the inconsistency: we can't consistently think that we don't have the one obligation if we think we clearly do have the other.

The asymmetry may also be put teleologically—in terms not of act but rather outcome evaluation. Then, one half asserts that, other things equal, the existence of the happy child does *not* make an outcome morally *better* than another. The other half asserts that, other things equal, the existence of the miserable child *does* make an outcome morally *worse* than another. The inconsistency? Again: we can draw no clear morally significant distinction between the one outcome's including the miserable child and the other outcome's excluding the happy child. And the contingency of the individual's existence does not strip the individual of moral status. Hence: if we say the outcome that includes the happy child isn't better than the one that doesn't, we can't consistently say that the outcome that includes the miserable child isn't worse than the outcome that doesn't.

Early efforts to preserve the asymmetry relied on the point that, if agents make the miserable child worse off by bringing that child into existence, there is an *existing victim*, a *flesh and blood person* who may be said to have a *complaint*. In contrast, if agents make the happy child worse off by *not* bringing that child into existence, there is no such victim, no flesh and blood person whom agents have in some way failed. This approach implicitly references to the intuition already noted in connection with the nonidentity problem, that is, the *person-affecting intuition*: no (existing or

future) victim, no wrong act; no (existing or future) person made worse off, no worse outcome.

Such a person-affecting account of the asymmetry seems tantalizing on its face. On closer analysis, though, Heyd argues that things are more complicated. The existing *miserable* child—if the miserable child exists; if *that* is how the choice is made; if *that* is the outcome under scrutiny—can indeed complain about having been made to exist. But if that's so then the existing *happy* child—if the happy child exists; if *that* is how the choice is made; if *that* is the outcome under scrutiny—may then, properly taking for granted that making a person happy has no lesser moral significance than does making a person miserable, "thank us for being born" (Heyd, p. 60). Put otherwise: if it's the *existence* of the complainer that puts the complaint on the moral radar, then so should the *existence* of the grateful put the expression of gratitude on the moral radar. And hence, again, an inconsistency: if we take the position that that it's wrong to bring the miserable child into existence or that the existence of the miserable child makes the outcome morally worse, then we can't at the same time take the position that it's permissible not to bring the happy child into existence or that the existence of the happy child does not make things better.

Heyd resolves the inconsistency by denying that agents are obligated not to bring the miserable child into existence. But most philosophers, e.g. Singer and Parfit, instead resolve the inconsistency by—contrary to intuition—denying that agents may permissibly leave the happy child out of existence.

My view is that the two halves of the asymmetry can be reconciled against each other. For that purpose, and consistent with a person-based approach, we can accept that a person's loss is morally significant if and only if the world where that loss is sustained is one where the person does or will exist and not otherwise. (I elsewhere call this principle the *loss distinction thesis* and, earlier on, *variabilism*.) Losses themselves are simple diminutions in a person's wellbeing level in one possible outcome

as compared against another. Thus we recognize that Hans sustains a loss in w3. But in contrast to the loss Meg sustains in w1, Hans's loss in w3 has no moral significance at all—counting neither against w3 nor, in a roundabout way, in favor of w1. The upshot: Meg's loss makes w1 worse than w2 and makes a1 wrong, while Hans's loss in w3 doesn't make w3 worse than w4 and doesn't make a3 wrong.

On a consequentialist approach, losses and gains are two sides of the same coin. We can thus also say that a gain is morally significant if and only if it reverses a morally significant loss. Analyzing the cases in terms of gains, we obtain results that simply repeat what we've just said. Meg's gain in w2 is morally significant; it counts in favor of w2 and against w1, making w2 better than w1 and making c1 wrong. Hans's gain in w4, in contrast, has no moral significance at all. w3 and w4 are equally good, and a3 and a4 are both permissible.

Having done that work, we can easily assign levels of personal good. Those are indicated in the graph. Adding these up, we confirm the results we've already noted.

[END OF DRAFT MS.]