Excerpts from a draft of *The Limits of Kindess* (OUP October 2013) for the Law and Philosophy Seminar at USC.

The session will be about two questions in procreative ethics:

1. The classic ‘same number non-identity problem’ -- Why should we create healthy people, rather than numerically distinct, mildly unhealthy people, when given the choice?

2. The ‘retrospective attitude problem’ -- Why, after we have created a mildly unhealthy person, should we be glad of having done so?

In Chapter 5 of my recent book I suggest an answer to the first question. That Chapter, pages 50 to 65 of this document, is the main reading for the session. My basic idea is that if we are decent and rational then we will choose to create healthy people. The conception of decency that supports this idea (decency requires minimal benevolence) is spelled out in Chapter 1 of the book, pages 24 to 33 of this document. The conception of rationality that supports this idea (prospectist rationality) is spelled out in Chapter 3 of the book, pages 34 to 49 of this document.

I also include the introduction to the book, pages 2 to 23 of this document, for people who want to get a sense of my larger project. That is very optional reading.

I presented some of this material at a colloquium at USC back in September 2010.

In a new, in-progress paper I suggest an answer to the second question. The paper is very rough, but I will forward it along.

Caspar
Introduction

1. Normative Ethics

Let’s begin with some basics. This book is about normative ethics. What is that?

There’s a natural distinction between claims like these:

“The square root of two is irrational.”

“The economy is in decline.”

“Proxima Centauri is our Sun’s closest big neighbor.”

“That really, really hurt.”

and claims like these:

“You ought to be nice to your grandmother.”

“Paris is beautiful.”

“Lying to little children to serve your own ends is wrong.”

“Television is bad for the soul.”

The first sort of claim pertains only to the way things are, the second sort of claim pertains, at least in part, to the way things ought to be. In philosophy-speak: the first sort of claim is descriptive, the second normative.¹

Some normative claims overtly concern morality. For example:

“Cyril is a dishonorable bounder.”

“The administration is irredeemably corrupt.”

“Harry wronged you by taking your hat.”

Others do not. For example:

¹ I should note (the first of many qualifications) that though the basic distinction is natural and easy to grasp, drawing a precise boundary between normative and descriptive claims is an exceptionally interesting and difficult job. We need not worry about that here.
“Eating your hair is a bad way to conserve energy.”

“On clay, Roger Federer ought to adjust his backhand to compensate for the higher bounce of the ball.”

Moralists are in the business of making normative claims of the first sort, moral-normative claims.²

Moral theorists are moralists with a systematic bent. They are unsatisfied with claims about particular things, like Cyril, or Harry’s action-of-taking-your-hat. They have grander ambitions. They want to make very general claims about kinds of things:

“All people of kind __ are dishonorable bounders.”

“All actions of kind __ wrong you.”

“All punishments of kind __ are unjust.”

Their idea is for these general claims to serve as the basis of a theory that will entail claims about particular things.

Normative ethicists are moral theorists with a special interest in individual people – in how they are, and in what they do.³ Normative ethicists are not directly concerned with when institutional structures are fair and just, when policies are discriminatory, when aspirations are noble, when customs are deviant and contrary to nature… and so on and so forth. They want a theory of the conditions under which people are good or bad, honorable or dishonorable, admirable or loathsome…etc. And they want a theory of the

² Again, I should note that, though the distinction between normative claims that have a moral flavor and normative claims that do not is natural enough, there is no uncontroversial way of drawing a precise boundary between the realm of the moral and the realm of the normative-non-moral. Again, we need not worry about that here.

³ This is how the term ‘normative ethics’ is standardly used by philosophers (when introducing the subject, philosophers standardly say that the central questions of normative ethics are ‘how should we live?’, ‘what should we do?’ ‘how should we be?’). There’s a worry that the term is misleading general – ‘the ethics of individual action’, or ‘the ethics of life’ might be better terms for what philosophers have in mind. But I am not going to try to change entrenched ways of talking here.
conditions under which people’s actions are right or wrong, permissible or impermissible, praiseworthy or blameworthy…etc. They are, not to mince words, high-minded busy-bodies. Their goal is to tell you how you ought, morally, to be and what you ought, morally, to do.

2. Why Normative Ethics is Hard

Suppose we try to join the high-minded busy-bodies, and set about building a theory of normative ethics. How hard can this be? Quite hard, it turns out. Let’s start by addressing the morality of action. Our first job is to find some general claims about the moral status of actions to build our theory around. But this is not a simple matter. Trusting our instincts, looking for general claims that have a plausible ring to them, will not help, because many general claims about the moral status of actions that have plausible ring to them turn to be inconsistent with one another. Normative ethicists have proven to be very adept at rooting out and exposing such inconsistencies. I will discuss here, and over the coming chapters, three celebrated examples of pairs of inconsistent, right-seeming normative claims. Here’s one pair:

**Betterness**
You do wrong only if things overall would have been better if you had acted in some other way.

**Violation**
It is wrong to inflict grievous physical harm on an unconsenting, innocent person who poses no threat to anybody, unless you can secure massively disproportionate benefits by doing so.
Betterness seems right. If I accuse you of doing something wrong, then I can’t very well say “…though, of course, things would have been no better if you had done anything else.” By accusing you I seem to be committing myself to the idea that things would have been better if you had done something else. And Violation sounds right too. Of course it is not okay to hurt unconsenting, innocent, unthreatening people.

But the claims are inconsistent. Consider:

Human Fuel
While piloting a steel, steam-engined boat across the bleak South Seas, you receive a distress call from Amy, who has been left to die of thirst on a nearby, bare island by cruel pirates. Knowing that you are the only boat in the area, you pick her up, and then receive another distress call from Brian and Celia, left to die of thirst on a not-so-nearby, bare island by more cruel pirates. You do not have enough coal to get to them, and no part of your steel boat will serve as fuel… but Amy is both large and dehydrated… Knock her on the head, shove her in the furnace and you will make it over there. And nobody but you will ever know.

If Violation is true then it would be wrong to use Amy-power to save Brian and Celia.
But if Betterness is true then, seemingly, it would not be wrong. If you knock Amy on the head, then one person will have been killed in a quick, relatively painless way. If you don’t, then two people will have been killed in a lingering, relatively painful way. It is no
better that two people be killed in a lingering, relatively painful way than that one person
be killed in a quick, relatively painless way.⁴

Here’s another celebrated example of a pair of inconsistent, right seeming claims:

\textit{Harm}

An action is wrong only if it harms something with morally significant interests.

\textit{Optimizing the Health of Your Child}

Whenever you have made it your business to conceive and bear a child, it is
wrong to choose that your child be unhealthy, rather than healthy, unless you
have strong reasons to do so.

\textit{Harm} seems right. If I argue that you have done the wrong thing, and you reply “Where
was the harm in that?” then I must answer your question or lose the argument. And
\textit{Optimizing the Health of Your Child} seems right too. Of course parents ought have some
concern for the health of their children. Of course they ought to act on that concern,
absent strong reasons to the contrary.

But the claims are inconsistent. Consider:

\textit{Not Postponing Conception}

Mary is recovering from German measles. Her doctor recommends that she
postpone her efforts to conceive a child for a couple of months. If she
conceives a child in the next couple of months then the child will, most likely,
have significant health problems. Mary has no strong reasons to conceive a

⁴ This problem goes by various names. It is sometimes referred to as ‘Scheffler’s Paradox’, after an
influential treatment of it in Scheffler (1992), sometimes referred to as ‘Foot’s Problem’ after an influential
treatment of it in Foot (1985). But philosophers were aware of some form or other of the problem well
before that. See, for example, McCloskey (1963). I will discuss the problem in more detail in Chapter
Seven.
child immediately, but she does have a mild preference for getting on with it. She gets on with it. Nine and a half months later baby Mariette is born, with significant health problems. This is not a disaster – Mary is a woman of means, so Mariette’s health problems do not impose a burden on wider society, and, on balance, Mariette has a rewarding life. But it is not great either – Mariette’s health problems are a chronic source of anxiety, pain and frustration to her.

If Optimizing the Health of Your Child is true then it would seem that Mary did wrong by conceiving her child immediately. But if Harm is true then it would seem that she did no wrong. If Mary had not conceived a baby immediately then she would, most likely, have conceived a baby some time later, as genetically different from the actual Mariette as typical non-twin siblings are genetically different. On any plausible view of essence (of the conditions under which particular people would or would not have existed), that child would not have been Mariette. On any plausible view of harm, you do not harm somebody whose life is, on balance, rewarding, by making it the case that he or she exists rather than not. So Mary didn’t harm Mariette, and if she didn’t harm Mariette, then who or what did she harm?

Here’s the third celebrated example of a pair of inconsistent, right-seeming claims:

**Sacrifice**

Whenever you are in a position to save the life of a child at a relatively small cost to yourself and no cost to others, you are morally obliged to do so.

---

5 This has been known as an instance of the ‘Non-Identity Problem’ since Parfit (1976), and Parfit (1983) Chapter 16. I will discuss it in detail in Chapter 5.
*Sackcloth and Ashes*

People living in affluent societies are not morally obliged to give away almost everything they have, for the sake of people in distant, poor societies.

*Sacrifice* sounds right. What kind of a monster would let a child die for the sake of his cufflinks? *Sackcloth and Ashes* sounds right too. Surely we do not have to give away almost everything we have! Can’t we hold onto at least a few small indulgences – a car when we could take the bus, freshly squeezed orange juice when we could get squash?

But the claims are inconsistent. Consider:

**Oxfam**

The charity Oxfam is soliciting donations for a program that will vaccinate impoverished children against disease. Craftily, the administrators of the program have arranged their finances in such a way that the marginal benefits of further donations are clear: For every $100 you give, around ten more children will be vaccinated. For every ten children vaccinated, around one of them will live through an epidemic of disease that would otherwise have killed them.

If *Sacrifice* is right then it would appear that you are obliged to give your first $100 to the program – by doing so you will save the life of a child at a small cost to yourself. And you are obliged to give your second $100 to the program for the same reason, and your third $100… right down to the point where the marginal cost to you of giving away a further $100 is relatively large. But by that point you will have given away almost
everything you have. So if Sacrifice is right then you are obliged to give away almost everything you have in this case, which contradicts Sackcloth and Ashes.\(^6\)

3. Reflective Equilibrium

The moral to draw from these examples is that not all general normative claims that seem right, at first blush, are right. How, then, are we to decide which to accept? Many normative ethicists would give this advice: “Building a theory of normative ethics is about working towards reflective equilibrium. Start by taking all the normative claims that sound right to you, whether they be very general claims about the nature of the good and so forth, or very particular claims about particular cases. Then test them against each other. Look for inconsistencies. If you find inconsistent claims then discard one or the other. When you make decisions about which claims to discard, do not dogmatically favor the more general over the more specific, or vice versa – consider the strength of your attachment to the respective claims, consider the unity, simplicity and explanatory power of the emerging theory. If you can find more general claims that will entail and explain disparate, surviving, more specific claims then adopt them. Repeat this process again and again, until you are left with a simple, consistent theory.”\(^7\)

I have never found this advice very helpful. I find it easy enough to identify claims that sound right to me, to test such claims against each other, and to find inconsistencies. But when it comes time to “discard one claim or the other” I often have no idea which way to go. Take the Human Fuel case, for example. Betterness and


\(^7\) The name ‘reflective equilibrium’ was coined by John Rawls in Rawls (1971). Rawls traced the idea back to Nelson Goodman’s discussion of methods for justifying rules of inductive logic in Goodman (1955).
Violation both seem right to me. The case shows they are inconsistent. How am I to proceed?

Attending, first, to the unity, simplicity and explanatory power of the emerging theory, I find that Betterness comes out ahead of Violation. Precisely explicating the virtues of unity, simplicity and explanatory power in a theory is a notoriously difficult problem. But any interesting explication of these virtues will cast act consequentialism, the theory that makes Betterness its centerpiece, in a flattering light. Act consequentialism can be stated in a sentence:

**Act Consequentialism**

An act is wrong if and only if some alternative to it has a better outcome.

When sharpened up a bit (we need to explain what alternatives and outcomes are) and combined with an appropriately precise axiology (a theory of what makes one outcome better than another) it nails down the moral status of all acts, performed by anyone, at any time, in any place. Violation, on the other hand, is a very local principle. It tells us about the moral status of certain acts that inflict grievous physical harm on unconsenting, innocent people, and nothing else. Some theorists who adopt Violation think of it as just one of a patchwork of independent moral principles. These principles together make up a moral theory that is far more complex and disunified than act consequentialism. Others think that Violation can be derived from principles that are (at least close to) as simple and unified as act consequentialism – variants of rule consequentialism, contractualism and Kantian rationalism. But I, for one, find these derivations less than convincing.
But unity, simplicity and explanatory power are not the be-all and end-all of everything in normative ethics. If they were then moral nihilism (for these purposes: the view that all actions are morally neutral) would be the best theory of all. The method of reflective equilibrium would have me weigh, also, the strength of my attachment to Betterness and Violation. How strongly am I attached to each?

Well, the first thing I must acknowledge is that, for me, Violation has a lot of pull. I expect it has a lot of pull for you too. Imagine ignoring it, and behaving as the act consequentialist would have you behave in the Human Fuel case. Imagine picking Amy up from the first island. Imagine her relief at seeing you. Imagine learning of the second island. Imagine shielding your calculations from Amy (better that she not know about them, really). Imagine creeping up behind her and smashing her head with a heavy spanner. Imagine smashing her head again and again to be sure that she is dead (very important, given what is about to happen). Imagine dragging her body across the deck and cramming it into the boiler. Imagine the temporary loss of power to your boat as the liquids in her body evaporate away. Imagine pungent new smells emanating from the boiler. Imagine the surge of power as her flesh begins to burn. Imagine arriving at the second island. Imagine Brian and Celia’s relief at seeing you. Imagine steaming away with that secret smoldering in the middle of your boat... I expect that, the more carefully you fill in the details of this story in your imagination, the more disgusting you will find it. And I expect that your disgust will have a particular flavor. It will not be the sort of disgust that comes with imagining doing something morally admirable but gross – like diving into a cesspit to save the life of a toddler, or spooning the brains of an injured
soldier back into his skull. It will be moral disgust. Your imaginary actions will seem to you, in a visceral way, to be wrong.

Fine. Now, what should you and I do with our visceral judgments?

On one way of thinking, we should discard them. All that matters in the Human Fuel case is that if you do the one thing then two people will be killed, while if you do the other thing then one person will be killed. Our visceral judgments about these sorts of cases are mistaken.

Advocates of this way of thinking owe us a debunking explanation of how we came to be mistaken. And they have many to hand. For example, they can say: “Doubtless the idea of knocking Amy on the head and throwing her in the boiler seems horrific to you, but this is because the immediate consequences of doing so are horrific, and we all have a tendency to focus on the immediate consequences of what we do. This tendency serves us well, most of the time, because most of the time we have far greater control over the immediate consequences of our actions than the distant consequences of our actions. But sometimes it leads us astray. It is leading you astray here. Immediate and distant consequences matter equally much.”

On another way of thinking, we should preserve the judgments. Obviously we do not want to demand of our incipient moral theory that it represent all of our visceral judgments about right and wrong as true. We want to leave space for correction. But this visceral judgment is non-negotiable. Its truth is radiant. Any claim with which it is inconsistent should be rejected.

Advocates of this way of thinking, too, owe us a debunking explanation of why we might be inclined to think otherwise, to think that killing Amy is in fact the thing to
do. And they, too, have many to hand. For example, they can say: “Look, what matters in this case is that Amy, Brian and Celia each have a right not to be murdered. Philosophers may appreciate that rights matter, and infer that the appropriate thing to do in this case is ensure that as few people as possible have their rights violated. Malaria is bad, so we should minimize malaria. Violation of rights is bad, so we should minimize violation of rights. But this is a subtle mistake. Rights are not like malaria. The appropriate way to respond to the fact that rights matter is not by minimizing the violation of rights, but rather by not violating rights.”

For each of the two points of view, I think I see it quite clearly, and I can work myself into a frame of mind where it seems like the right one. But which is the right one? I find it very hard to judge. I know, of course, what I would do if I found myself in the Human Fuel case. I would leave Amy alone. I doubt that I would even allow myself to acknowledge that throwing her into the boiler was an option. It would be one of those thoughts that flits around behind the camera of my mind, too shameful and disturbing to be released into view. But would that reflect some deep practical wisdom on my part, or would it reflect moral cowardice? I find it very hard to judge.

Is this inability to judge a weakness on my part? Maybe so. Certainly, many of my friends and colleagues have no difficulty with the question. They whole-heartedly embrace one of the views and renounce the other. They think it obvious that we should all do the same. But I take some consolation in the fact that they do not all go the same way. Roughly half of them embrace some form of act consequentialism. Roughly half of

---

8 See Pettit (1991) for a very clear version of this point.
them embrace some form of deontology – some view that yields the result that Amy should be allowed to live.

In my experience, feelings run pretty high in this domain. Many act consequentialists tend to think of deontologists as wooly-headed weaklings, as people who would rather the obscure the landscape of normative ethics with enigmatic musings about Kant than face up to the fact that, sometimes, contrary to ‘intuition’, they must dirty their hands. Many deontologists tend to think of act consequentialists as accountants-gone-wild, as people who have entirely lost touch with their moral sense, and come to care only about the books. (Nine years ago, when I first arrived at MIT, I taught a graduate seminar on normative ethics. Judy Thomson, who I enormously respect and admire, attended the seminar, and came to suspect that I had an unseemly attraction to act consequentialism. One day, after class, she pressed me on cases like the Human Fuel case. When I confessed to feeling the pull of the act consequentialist way of thinking about such cases she threw her hands in the air, gave me a special sort of look, proclaimed my condition ‘terribly sad’\(^9\), and walked away. The look made a big impression on me. It wasn’t the fond, indulgent sort of look that you give your wayward young nephew when he tells you that he is thinking of the joining the socialist party. It was the sort of look that you give your wayward young nephew when he tells you that he is thinking of joining the Klan.)

It is a curious phenomenon. These are very intelligent people. They are fully aware of the arguments on both sides of the issue. Ask the act consequentialists to make a

\(^9\) But please do not infer that Judy Thomson’s grounds for rejecting act consequentialism come down only to this feeling of terrible sadness. She has a very sophisticated account of where and why act consequentialists go wrong. I will briefly summarize it in Chapter 4 section 6.
case for deontology, for pedagogical purposes, and they will say everything the deontologists say, with all the same passion, emphasis and conviction. Ask the deontologists to make a case for act consequentialism, for pedagogical purposes, and they will say everything the act consequentialists say, with all the same passion, emphasis and conviction. It does not appear as if the members of either group are missing something – as if there is a decisive consideration that has passed beneath their attention. And they all know this. Yet some psychological mechanism causes roughly half of them to break one way and roughly half of them to break the other way.\textsuperscript{10}

4. A Foundational Approach

If you, like me, don’t break either way, if you share my ambivalence about these sorts of questions, then you might be interested in a different approach. You might hope to set the ‘prima facie plausible’ normative claims aside and base your theory on firmer foundations. You might hope to fix on some general normative principles that you regard as non-negotiably true. You might hope to take these principles as axioms and derive a

\textsuperscript{10} An untutored outsider might feel that this reflects badly on what normative ethicists are doing. The three problems that I have drawn attention to here are among the most basic problems in normative ethics. They are routinely taught in introductory classes on the subject. The fact that there is no consensus on how to resolve them, and a firm consensus that our present methods of enquiry will not yield a consensus on how to resolve them, might suggest, to the untutored outsider, a weakness in those methods. Philosophers tend to be wary of indulging such feelings (with some reason – in philosophy, deadlock is everywhere). A good part of our early education consists in stripping them away. David Lewis used to say: “Debates may deadlock. But it does not follow that they are not worth pursuing, or that there is no fact of the matter about who is right.” Back when I was young and impressionable, this struck me as deep wisdom, but now I am not so sure that the moral we were expected to draw from it (that, in philosophy, there’s a kind of virtue in sticking to your guns in the face of implacable opposition) is a good one. Often, when philosophical debates deadlock, it seems to me irresponsible to come down firmly on one side or other. Often it seems to me that coming down firmly on one side involves willfully ignoring the powerful considerations that move your opponents.
normative theory from them. You might hope that the resulting theory would be as solid and pure as an axiomatization of number theory.\textsuperscript{11}

What sort of ‘non-negotiable’ principles would do the job? A tempting place to start is with principles concerning rationality – principles concerning what it is rational to believe, desire and do.

This is not an original idea. For almost as long as there have been philosophers, there have been philosophers trying to base ethics on rationality, broadly understood. Plato at least experimented with the thought that morality was about enlightened self-interest. The ancient stoics held as their ideal a life in accordance with reason, free of internal conflict. Kant claimed that moral requirements derived from his categorical imperative, and that his categorical imperative was a requirement of rationality. Generation after generation of neo-Kantians have argued that rationality requires us to be impartial in some way, and this requirement is the source of moral obligation. Most, following the letter of Kant, have argued that the resulting obligations are deontological. Some, like my unrelated namesake R.M. Hare, have argued that the resulting obligations are consequentialist.

Airily sweeping my hand across thousands of years of intellectual history, I say now that none of these projects does the job we want done here. There’s a reason why. To derive substantive moral principles from principles of practical rationality you need some very rich principles of practical rationality. But the principles of practical

\textsuperscript{11} I should mention that some philosophers might say that this, too, is an application of the reflective equilibrium method, because they characterize the method more broadly than I have done here. Any form of rational inquiry whose goal is a state of reflective equilibrium, a state in which your more general beliefs and more specific beliefs are no longer in conflict (which is to say \textit{pretty much any form of rational inquiry}) counts as an application of the reflective equilibrium method. That’s fine by me. But then the advice ‘pursue reflective equilibrium’ is only marginally more helpful than the advice ‘think harder about the problem’.
rationality that have the non-negotiable flavor that would make them suitable as foundations for a robust normative theory are relatively impoverished.

When we think of non-negotiable principles of practical rationality the examples that spring to mind have to do with internal coherence: if you are rational then your desires are internally coherent (roughly: you do not want one thing and at the same time want another – I will unpack this precisely later) and your behavior is coherent with your desires (roughly: you do not want one thing and do another – again, I will unpack this precisely later). But it is easy to be internally coherent. Internally coherent people may be good people, who want good things and behave in good ways, or evil people, who want evil things and behave in evil ways, or just bizarre people, who want bizarre things and behave in bizarre ways. The history of philosophy is full of examples that illustrate this point. The most famous are due to David Hume. If we take ‘reason’ to place nothing more than coherency constraints upon us\(^\text{12}\) then we must agree with him that

\[\text{’Tis not contrary to reason to prefer the destruction of the world to the scratching of my finger. ’Tis not contrary to reason for me to choose my own total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. ’Tis as little contrary to reason to prefer even my own acknowledged lesser good to my greater, and have a more ardent affection for the former than the latter.}\(^\text{13}\)

\(^{12}\) Did Hume himself think that reason placed coherency constraints upon us? He did write that someone who fails to choose the proper means to his end makes an ‘error’ – which might appear to suggest that he thought of the instrumental principle (which says, roughly: \textit{take the known means to your desired ends} – a sort of coherency constraint) as a requirement of rationality. But there are grounds for thinking that the appearance is misleading, that Hume had in mind a person whose error was one of falsely believing something to be a means to his end (see Korsgaard 2008).

\(^{13}\) From Book 2, Part 3, section 3 of Hume (1740). The examples are under-described in odd ways. Does the first fellow prefer the destruction of his finger to the scratching of his finger? If not, is his finger not part of the world? – But the general moral is clear.
It does not follow from my being internally coherent that I will behave in one way or another when placed in a morally portentous situation. I may be an internally coherent saint or an internally coherent psychopath. If we are to derive a substantive, interesting theory of normative ethics from rock-solid axioms then we will need to supplement principles of internal coherence with something richer.

What might do the trick? One strategy might be to add axioms concerning about the *content* of rational desires – about what it is rational to desire. Some philosophers have claimed that rationality places significant constraints on what we desire. Derek Parfit, for example, has said\(^{14}\) we have reasons for desiring various things, and that we are more or less rational to the extent that our desires are more or less strongly supported by reasons. Someone whose desires are wildly out of step with reasons (his canonic example is a man who has a healthy desire that he not suffer pain, except on Tuesdays – a man who would rather that he suffer any amount on pain on a Tuesday than any amount of pain on any other day of the week) is, by any standard, irrational. So we might add an axiom that says that if you are rational then your desires are not wildly out of step with reasons.

Maybe this claim has the non-negotiable flavor that we want of the axioms of our theory. But to derive any interesting conclusions about what rational people do in morally portentous situations, we will need to supplement it further, with some specific claims about which kinds of desires are, and which kinds of desires are not, out of step with reasons. (Jane kills one person to prevent two from being killed, because she cares more about *there being less killing* than about *her not killing*. John refuses to kill one person to

---

\(^{14}\) In Part One of Parfit (2011a).
prevent two from being killed, because he cares more about *his not killing* than about *there being less killing*. If we are to derive any interesting conclusions about their respective actions, we have to say that one of John or Jane has desires out of step with reasons.) These sorts of claims certainly will not have the non-negotiable flavor that we want of the axioms of our theory.

In light of considerations like this, many contemporary philosophers think that the project of deriving a substantial theory of normative ethics from self-obvious axioms is hopeless. Normative ethics is not about cranking out results. It is an altogether subtler business. To do normative ethics well you must do reflective equilibrium well. And to do reflective equilibrium well you must have tact, experience, and sensitivity to the delicate nuances of moral life. This is why pimply adolescents are very good at number theory, but very bad at normative ethics.

I am no longer a pimply adolescent. But I find that the little tact, experience and sensitivity to the delicate nuances of moral life that I have gained since my pimply adolescence leave me ill-equipped to settle any interesting questions in normative ethics by balancing my intuitions about principles against my intuitions about cases in the subtle ways these philosophers recommend. So, in this book, I want to pursue a more foundational, bottom-up approach further. I think that we can make significant progress in normative ethics by supplementing some very minimal assumptions about rationality with some very minimal assumptions about moral decency.

5. Moving Forward

In Chapter One I spell out my first assumption about moral decency. It amounts to
this: If you are decent then you are at least minimally benevolent towards other people. *When absolutely all other things are equal*, at least, you would rather that other people be better off rather than worse off.

In Chapters Two and Three I spell out my first assumption about rationality: If you are practically rational then, when you don’t know what will happen if you do one thing or another, your decisions are guided by the prospects associated with the acts open to you. This is a very intuitive idea, though tricky to put in a precise way, as we will see.

In Chapters Four to Six I put these assumptions to work in normative ethics. Chapter Four is about saving people from harm. There has been a great (albeit, to outsiders, rather mysterious) controversy in the normative ethics literature over whether and (if so) why, given the choice, we are obliged to save more people, rather than fewer people, from similar harms, and to save a multitude of people from a small harm rather than one person from a large harm. But it follows from the minimal assumptions that, in some circumstances at least (circumstances in which you don’t know who you will save by doing one thing or another), if you are decent and rational then you will save more people, rather than fewer, from similar harms, and you will save the multitude from the small harm rather than the one from the large harm.

Chapter Five is about cases like *Not Postponing Conception*, cases that raise ‘the non-identity problem’. It follows from our assumptions that, if Mary were decent and rational, then she would not have conceived unhealthy Mariette. It does not follow that she has done something *wrong*. But I will suggest that the two notions *being such that somebody decent and rational would not do it*, and *being wrong*, are closely connected.
Chapter Six is about cases like *Human Fuel*, cases that raise the problem of whether it is okay to kill-to-prevent-two-killings. I will argue that, by framing the issue in terms of what a minimally decent and rational person will do in these cases, we add significant force to a traditional objection to the deontologist’s treatment of them: the so-called ‘dirty hands’ objection to deontology.

Thus far we have only made progress in cases in which you do not know who you are in a position to harm or benefit by doing one thing or another. In Part II of the book, Chapters Seven to Eleven, I will extend the treatment to cover cases in which you do know who you are in position to harm or benefit by doing one thing or another. To do it I will need fancier tools than before.

In Chapter Seven I spell out an assumption about people and their essences. Each of us could have been ever-so-slightly-different along any natural dimension of sameness and difference. You could have been a millimeter taller than you actually are. I could have been conceived a second before I was actually conceived. Barak Obama could have been slightly more irascible than he actually is. Our essences are not perfectly fragile.

In Chapter Eight I spell out a further, quiet assumption about rationality. If you are rational then your desires are coherent – which means, at least, that your preferences between maximal states of affairs (fully specific ways for everything to be) are transitive.

In Chapters Nine to Eleven I put these new assumptions to work. In Chapter Nine I argue that it follows that, if decency obliges us to prefer, in some circumstances, that some people be better off rather than worse off, then decency and rationality together oblige us to prefer, in some circumstances, that some people be better off rather than other people better off. I call this the *Morphing Argument*. I explore its consequences for
the problems we have looked at so far, problems involving who to save, the non-identity problem, and the problem of whether to kill-to-prevent-killing, in Chapter Ten.

Chapter Eleven is about the limits of good-will towards others. One surprising consequence of the morphing argument is that it is impossible to be both rational and minimally benevolent towards everyone. Rationality itself places limits on how good-willed we can be.

Which brings us to the third major problem that I discussed in the Introduction – our moral obligations towards distant strangers. In Chapters Twelve and Thirteen I argue that the rational requirement that preferences be transitive and the moral requirement that we be at least minimally benevolent towards (some) strangers together place great pressure on us to attend to the plight of needy, distant strangers. It is much harder than we ordinarily think to be both decent and rational. In Chapter Fourteen I moderate the claim somewhat. Rationality and decency do not demand of us that we be equally devoted to everybody. Being rational involves being committed in a certain way to particular people and their interests.

6. Two Goals

Part of what I hope to do here is to shed light on some interesting problems in normative ethics. Another part of what I want to do is to show off the benefits of approaching the subject in this way – by thinking about rationality and minimal benevolence. In typical introductory classes in moral philosophy we teach our undergraduates that there are three approaches to normative ethics. There’s the consequentialist approach (usually traced back to Jeremy Bentham – with a nod to the
ancient Epicureans), which has it that acting morally is about bringing about the best states of affairs we can. There’s the deontological approach (usually traced back to Kant, with a nod to contractualist and rights-based alternatives), which has us focus on the character of the act itself, with an emphasis on whether it is universalizable – on how we would feel about everybody acting this way. And there’s the virtue-based approach (usually traced back to Aristotle), which has us focus on how the act reflects on the moral character of the agent – is it the sort of thing that a perfectly virtuous (honest, kind, loving, courageous…etc.) person would do? The approach I am exploring here does not fit neatly into any of these traditions. I think this is a good thing, as you will see.
Chapter 1: The Good Will

Here’s my first claim: being morally decent involves adopting an attitude of good will towards other people, which involves taking their interests into account, and wishing them well.

This will, I expect, strike you as an obvious truth. Indeed it may strike you as such an obvious truth, such an anodyne, saccharine platitude, that you wonder why an author would bother to draw attention to it. In writing it, I am reminded of passages in It Takes a Village: And Other Lessons Children Teach Us, in which Hillary Clinton sincerely intones that:

“\text{In addition to being read to, children love to be told stories.}^{16}\text{ and} \text{Brisk walking, hiking and bicycling are all good exercise, and are great ways to spend time together as well.}^{17}\text{ and} \text{The lessons men and women have learned over thousands of years are available to anyone, in the form of fables, stories, poems, plays, proverbs, and scriptures that have stood the test of time.}^{18,19}

She is not wrong. These things are true. But of all the many, many true things that she could have chosen to draw her reader’s attention to, they are among the blandest and least interesting.

But we are doing philosophy now, and philosophers are suspicious people. If Hillary Clinton had published It Takes a Village in a philosophy journal then there would have been objections:

\begin{footnotesize}
16 ibid. p. 94
17 ibid. p. 108
18 ibid. p. 137
19 Credit goes to Martin Amis for drawing attention to some of these passages in his (1996) review of the book.
\end{footnotesize}
“So you claim that bicycling is good exercise. What about bicycling downhill, when we would otherwise have to walk? That doesn’t sound like very good exercise.”

“So you claim that brisk walking is a great way to spend time together. What about when we are briskly walking with someone very unfit, who struggles to keep up, and resents us for it? That doesn’t sound like a great way to spend time together.”

This would not just be down to temperamental pedantry on the part of the objectors. They have good reason to be suspicious of prima facie obvious truths. In philosophy there’s a grand tradition of deriving radical, surprising, wild claims from prima facie obvious truths. So, for example: the philosopher makes a few seemingly harmless claims about material objects. You nod along. Then, Boom! He shows that by nodding along you have committed yourself to denying the existence of composite inanimate objects. There are no tables and chairs. The philosopher makes a few seemingly harmless claims about morality and impartiality. You nod along. Then, Boom! She shows that by nodding along you are committing yourself to the existence of a number, $n$, such that, if the opportunity arises, you ought to kill your mother to cause $n$ rabbits to experience mild, brief pleasure.

The lesson is that when a philosopher asks you to embrace a prima facie obvious truth it is worth stepping back a moment and inspecting the object of your embrace. In this case I have said that being moral involves adopting an attitude of good will towards

---

20 (See Van Inwagen 1990).
21 From a philosophy-joke (I don’t know its origins):

Psychopath: ‘If you kill your mother then I will cause momentary sexual gratification to sweep through a shed-full of rabbits.’

Hedonic Utilitarian: ‘How big is the shed?’
others, which involves taking their interests into account, and wishing them well. It is worth your asking some questions. What exactly do I mean by ‘…taking their interests into account and wanting, wishing them well’? What exactly do I mean by ‘involves’? And who, for that matter, are these ‘others’? I will address these questions in order.

1.2 What is it to Take Another Person’s Interests into Account, and Wish Them Well?

Sometimes I turn my attention to two states of affairs (think of these as ways for things to be), and wonder which I would rather came about. So, for example, if you were to float the idea of my submitting myself to a severely calorie-restricted diet, I might think about the consequences of my going either way:

**Caloric Restriction**

- **In Life-Diet:** I restrict my long-term caloric intake – which gives me a longer, but less vigorous life.
- **In No-Diet:** I eat away – which gives me a shorter, but more vigorous life.

Sometimes I take certain considerations to be reasons to prefer one state of affairs over the other. So, for example, I might take the consideration ‘in Life-Diet I live longer’ to be a reason to prefer *Life-Diet*. And I might also take the consideration ‘in No-Diet I have more vigor’ to be a reason to prefer *No-Diet*.

My first claim here concerns the considerations that a morally decent person takes to be reasons to prefer states of affairs:
**Minimal Consideration**

Being morally decent involves being *minimally considerate* towards others – I am minimally considerate towards you when, for any states of affairs S, S*, I take the consideration ‘you are better off in S than in S*’ to be a reason to favor S over S*.

My second claim directly concerns a morally decent person’s preferences:

**Minimal Benevolence**

Being morally decent involves being minimally benevolent towards others – I am minimally benevolent towards you when, for any states of affairs S, S* such that you are better off in S than in S*, and such that I take myself to have no reason to prefer S*, I prefer S.

These are very weak claims. You probably believe something much stronger. You probably think that being decent not only involves taking other peoples’ interests into account, but also taking other peoples’ interests to be decisive in some cases when there are other things at stake. Consider:

**The Fate of the FA Cup**

<table>
<thead>
<tr>
<th>In Lose:</th>
<th>Arsenal lose the FA cup.</th>
</tr>
</thead>
<tbody>
<tr>
<td>In Win:</td>
<td>Arsenal win the FA cup after the Manchester United keeper accidentally, acrobatically, hangs himself on the netting of his goal.</td>
</tr>
</tbody>
</table>

I feel for the Manchester United Keeper, so I take the consideration ‘in Win the fellow dies a gruesome, public death’ to be a reason to favor *Lose*. But I also support Arsenal, so I take the consideration ‘In *Lose* Arsenal lose’ to be a reason to favor *Win*. You probably think that my being moral involves my taking the former consideration to be decisive. And you probably think that my being moral involves my preferring *Lose* to *Win*. 
Well and good, but that is an easy case. There are difficult cases in which serious, well-intentioned people will disagree about whether being decent involves taking other people’s interests to be decisive when there are other things at stake. Consider:

**Inequality**
- **In Gold:** Gardeners working for Mitt Romney discover rich seams of gold beneath the grounds of his New Hampshire property.
- **In No-Gold:** Gardeners working for Mitt Romney do not discover rich seams of gold beneath the grounds of his Aspen estate.

Some people will say that my being decent involves my preferring *Gold* to *No-Gold*. Mitt is better off in *Gold* than in *No-Gold*, end of story. Others will say that my being decent involves my caring about material inequality. It is at least not indecent of me to take the consideration ‘in *Gold*, material inequality is yet more egregious’ to be decisive. It is at least not indecent of me to prefer *No-Gold*.

Or consider:

**Property**
- **In Smoking:** Jake consumes too many cigarettes. He is unhealthy and unhappy.
- **In No Smoking:** I steal all of Jake’s cigarettes. He is healthier and happier.

Some people will say that my being decent involves my preferring *No Smoking* to *Smoking*. Jake is better off in *No Smoking* than in *Smoking*, that’s what matters. Others will say that my being decent involves my caring about property rights. It is at least not indecent of me to take the consideration ‘in *No Smoking* I violate Jake’s property rights’ to be decisive. It is at least not indecent of me to prefer *Smoking*.

Or consider:
Murder
In *Injustice*: Oliver gets away with the murder of a friend-less, relative-less person. Chastened by the experience, he lives a blameless life from then on.
In *Justice*: Oliver does not get away with the murder. He spends a long time in jail.

Some people will say that my being decent involves my preferring *Injustice* to *Justice*. Oliver is better off in *Injustice* than in *Justice*, end of story. Others will say that my being decent involves my caring about just deserts. It is at least not indecent of me to take the consideration ‘in *Justice* Oliver gets the punishment he deserves’ to be decisive. It is at least not indecent of me to prefer *Justice*.

And there are further difficult cases in which, though most people will agree that my being the embodiment of moral perfection involves my taking other peoples’ interests to be decisive, many serious, well-intentioned people will disagree about whether I stray far from basic moral decency by failing to do so. Recall the case from the Introduction:

Charity
In *Give*: I give a thousand dollars to charity, and some distant children are vaccinated against common diseases.
In *Keep*: I do not give any money to charity, and the same distant children are not vaccinated against common diseases, and some of them later contract these diseases and die.

Some people will say that I stray far, far from basic moral decency by failing to prefer, all things considered, *Give*. Some distant children are much, much better off in *Give* than in *Keep*. Others will say that I do not stray far from basic moral decency if I take the consideration ‘I am richer in *Keep*’ to be decisive, and prefer *Keep*. 
Well and good. I will broach the thorny issue of how to weigh considerations having to do with other people’s interests against other considerations later on. But for the moment let’s just assume that being moral involves being at least minimally considerate of others and minimally benevolent towards others. Let’s see how far we can go with that. For clarity, when we have to make a stronger assumption to make progress, I will FLAG it.

1.3 What is it to Say that Being Moral Involves Being Good-Willed Toward Others?

Some philosophers say that being moral involves exemplifying certain virtues – honesty, integrity, courage and so forth. Others say that being moral involves bringing about the best available states of affairs. Others say that being moral involves respecting rights. Yet others say that being moral involves giving people what they are owed, under the terms of real or hypothetical contracts.

I do not intend my claim to be in competition with any of these. Maybe being moral does involve respecting rights, giving people what they are owed, bringing about good states of affairs… and so forth. My claim is just that it also involves taking other people’s interests into account and wishing them well. There is a failing in someone who is honest and courageous, who respects rights, who gives people what they are owed, who brings about good states of affairs, but is indifferent towards the well-being of others, or positively malevolent towards others. And it is a moral failing.

Let me illustrate the idea with a story.
Bertha and Ben

Bertha and Ben are model citizens in almost every way. They are infallibly courteous towards others. They keep their promises. They do not lie. They respect rights. They run their lives by sophisticated variants of the golden rule: do unto others as you would have them do unto you. They think it is important for them to behave in chivalrous ways, so they help old ladies across the street whenever they can. They think it is important for them to behave in charitable ways, so they give large sums of money to the poor. They are admired by all who know them for being upstanding, tip-top people. But, in spite of all their good deeds, Bertha and Ben are not good-willed. Bertha is indifferent to the plight of others, and Ben positively malevolent towards others. One morning Bertha and Ben read that, in northern India, a small boy has fallen down a deep, narrow well. Rescuers are frantically digging down in an effort to save him. The news leaves Bertha quite cold. She cares no more about whether the boy lives or dies than you care about whether a particular drop of rain falling from the Pacific sky evaporates before it hits sea water. The news does not leave Ben cold. He hopes that the walls of the well will collapse, and that the boy will be trapped in an air pocket, in which he will suffocate slowly, screaming for his mother.

I claim that both Bertha and Ben, though admirable in many respects, are not, morally speaking, all they should be. I have no real argument for this claim. But arguments have to stop somewhere. This seems to me as good a place as any.

1.4 Who are these ‘Others’ Towards Whom We Must be Good-Willed?

I want to leave this last question open for the moment, largely because I want my claim in this chapter to be planted squarely in the realm of the obvious, and I think it far
from obvious just how far our considerateness and benevolence must extend. Take me. I am considerate of, and benevolent towards, my own children, and towards strangers from Northern India, but good will has its limits. It does not extend to past strangers, for example. When I learn a little about their lives, and am preparing to learn more, I find that I do not always wish to learn that things went well for them.

So, confronted with this fragment of American history, a letter from Alexander Hamilton, founding father, to his wife, Elizabeth Hamilton, on the subject of their son, Philip Hamilton, then fifteen years old:

To Elizabeth Hamilton
Rye 30 Miles from
New York
Tuesday Even [September 12, 1797]

I am arrived here My Dear Eliza in good health but very anxious about my Dear Philip. I pray heaven to restore him and in every event to support you. If his fever should appear likely to be obstinate, urge the Physician to consider well the propriety of trying the cold bath—I expect it will, if it continues assume a nervous type and in this case I believe the cold bath will be the most efficacious remedy—but still do not attempt it without the approbation of the Physician. Also my Betsey how much do I regret to be separated from you at such a juncture. When will the time come that shall be exempt from the necessity of leaving my dear family? God bless my beloved and all My Dear Children.

AH

I find that I do not particularly wish that the cold baths did their job and little Philip recovered from his infection. And that is not because I think there is anything else at stake. It is not that I loathe Hamilton and wish ill of his kin. It is not that I think that I think that some kind of sweet justice would have been served if Hamilton were punished

---

22 Printed in Hamilton (Allan) (1910).
for his absentee parenting. It is not that I think his life story would be more poignant if he suffered this loss. No, the whole affair just leaves me cold.

Is this a moral failing on my part? I am inclined to think that it is not. Maybe you disagree.\textsuperscript{23} Maybe you think it a failing because you think that being moral involves being at least minimally considerate of, and benevolent towards, \textit{everyone, always}. This is a noble view, but it cannot be right. I will argue, in Chapter 11 of this Book, that it is impossible to be both rational and good willed towards everyone, always.

\textsuperscript{23} In which case you will be relieved to hear that Philip did recover, though he was killed four years later, in a duel, pre-figuring his father’s death, also in a duel, two years after that.
Chapter 3: Rational Responses to Sweetening Insensitive Preferences

It often happens that there are items A, A+, B, B+ such that we prefer A+ to A, B+ to B, but we do not prefer A or A+ to B or B+. In the last chapter I drew attention to three cases in which I might have patterns of preference like this: The Fire, The Dinner and Small Differences, Fully Transparent Inter-Personal Conflict.

What is it rational for me to do in such cases? Classical expected utility theory will not help us answer this question. To apply classical expected utility theory we must associate a utility function, $U$, with my preferences such that, for all outcomes $x, y$, $U(x) > U(y)$ iff I prefer $x$ to $y$. Such a function exists only if my preferences between outcomes are negatively transitive (which means, recall, that for any states of affairs $x, y, z$, if I have no preference between $x$ and $y$, and no preference between $y$ and $z$, then I have no preference between $x$ and $z$). My preferences are negatively intransitive (I have no preference, recall, between A and B, and no preference between B and A+, but I prefer A+ to A).

“So much the worse for the question”, a classical expected utility theorist might say. “There’s no sense in asking what it is rational for you to do while your preferences are negatively intransitive. If you want guidance from the theory of practical rationality, reflect a bit, render your preferences negatively transitive, and then come back to us.”

This is an unhelpful response. I looked to the theory of practical rationality for guidance. It gave me none. It will guide me if I render my preferences negatively transitive, but I have no inclination to do that. And even if I did have an inclination to do it, doing it would involve acquiring or dropping preferences. It is not so easy to acquire or drop preferences at will.
A more constructive response is to extend the standard theory of rational decision under conditions of uncertainty to cover situations in which we have negatively intransitive preferences. But there’s a curious problem that comes up as soon as we try to do this. I will describe this problem, and two solutions to it, in this chapter.

3.2 Opaque Sweetening

Let’s consider a case that is less emotionally charged, a case that does not involve desperate people starving or thirsting to death on desolate islands. Let’s suppose that you have two items I want. Call them ‘A’ and ‘B’. Let’s suppose, also, that I would rather have A-and-a-dollar than just A, and rather have B-and-a-dollar than just B, but that I have no preference between having A and having B, between having A and having B-and-a-dollar, between having A-and-a-dollar and having B, or between having A-and-a-dollar and having B-and-a-dollar. And let’s suppose that we play a sort of game:

Two Opaque Boxes
You show me items A and B, a dollar, a coin, and two opaque boxes. You toss the coin and, governed by the toss (heads—left, tails—right), place item A in one box and item B in the other. I don’t see which item went where. You toss the coin again and, governed by the toss, place the dollar inside the right box. I see that, which leaves me with credence 0.5 that things are like so:

Left Box | A  | Right Box | B+$1
---|---|---|---

and credence 0.5 that things are like so:

Left Box | B  | Right Box | A+$1
---|---|---|---

Then you invite me to walk away with one of the boxes.
3.3 Why You Might Think I Should Take the Sweetened Option

Given what I know and prefer, what is it rationally permissible for me to do in this case? Here are two seemingly powerful arguments to the conclusion that it is rationally impermissible for me to take the left, unsweetened box:

*Argument 1: I Have No Reason to Take the Left, Rather than the Right, Box*

Imagine I take out a sheet of paper and try to list the pros and cons of taking the left and right boxes. What can I write in the Pro-Right-Rather-Than-Left column? Here’s one thing I can write:

**Pro-Right:** If I take the right box then I will get that extra dollar, but if I take the left box then I won’t.

What can I write in the Pro-Left-Rather-than-Right column? Nothing. This will not do:

**Pro-Left:** If I take the left box then I will get something I want.

If I take the right box then I will get something I want too. Nor will this do:

**Pro-Left:** If I take the left box, then there’s a .5 chance I will get A.

If I take the right box then there’s a .5 chance I will get A too. There’s no consideration that I can produce in favor of taking the left box rather than the right box. In philosophy-speak: I have a reason to take the right box rather than the left box, no reason to take the left box rather than the right box.

Now here’s a claim about practical rationality and reasons:
Rational Permissibility Tracks Reasons

If you are rational and two options are open to you, and you have a reason to take the one-rather-than-the-other, and no reason to take the other-rather-than-the-one, then it is rationally impermissible to take the other.

Being rational involves being guided by your reasons. When you have a reason to do one thing, and no reason to do anything else, then your reasons are guiding you in only one direction.

It follows that it is rationally impermissible for me to take the left box.

Argument 2. I Will Improve my Prospects by Taking the Right Box

Let the prospect associated with an option be, roughly, a representation of how likely I think it that one thing or another will happen if I take the option. To be precise, let it be the set of pairs \(<c,o>\) such that \(o\) is an outcome that might, for all I know, come about if I take the option, and \(c\) is my credence that the outcome will come about if I take the option.24 Here’s a claim about prospects and rational permissibility:

Prospects Determine Permissibility

Facts about what it is rationally permissible for me to do are determined by facts about the prospects associated with the options available to me.

---

24 If we wish to avoid committing ourselves to the denial of causal decision theory, then we should understand my ‘credence that the outcome will come about if I take the option’ in particular way. It is not just my credence in the outcome, conditional on my taking the option. Rather, my credence in outcome \(o\), relative to option \(a\), is \(\sum_d P(d).P(o/ad)\), where \(d\) is a variable ranging over dependency hypotheses, ‘\(P(d)\)’ refers to my credence in dependency hypothesis \(d\), and ‘\(P(o/ad)\)’ refers to my credence in outcome \(o\), conditional on my taking option \(a\) and dependency hypothesis \(d\) being true. But this will not make a difference in any of the cases I discuss here.
What it is rationally permissible for me to do depends only on the things I think might happen if I take the options open to me, and how likely I think them to happen.

Now consider another game:

**One Opaque Box**

You show me items A and B, a dollar, a coin, and one opaque box. You toss the coin and, governed by the toss, place item A or item B in the box. I don’t see which. Then you invite me to walk away with the box and the dollar, or just the box.

Obviously I have to accept the dollar in this case. But the prospects associated with the options available to me in this case are the same as the prospects associated with the options available to me in the Two Opaque Boxes case. In this case, the prospect associated with my taking the box alone is \{<0.5, A>, <0.5, B>\} (which is to say that I think it 0.5 likely that I will end up with A, 0.5 likely that I will end up with B, if I take the box alone), and the prospect associated with my taking the box and the dollar is \{<0.5, A+>, <0.5, B+>\}. In the Two Opaque Boxes case the prospect associated with my taking the left box is \{<0.5, A>, <0.5, B>\}, and the prospect associated with my taking the right box is \{<0.5, A+>, <0.5, B+>\}. So, by *Prospects Determine Permissibility*, in the **Two Opaque Boxes** case it is rationally impermissible to take the left box.

---

25 “But they are not exactly the same.” you might say. “In the first case, the prospect associated with my taking the left box is \{<0.5, I get A and could have gotten B+>, <0.5, I get B and could have gotten A+>\}. In the second case the prospect associated with my taking the box alone is \{<0.5, I get A and could have gotten A+>, <0.5, I get B and could have gotten B+>\}. Different prospects.” True, and if, in addition to caring about what I get, I care about whether what I get is preferable to what I leave on the table, then I have reason to treat this difference as significant. But if I don’t care about whether what I get is preferable to what I leave on the table, then I have no reason to treat this difference as significant.
3.4 Why You Might Think it is not the Case That I Should Take the Sweetened Option

Is that the end of the matter – I have to take the right box? Maybe not. Here are two seemingly powerful arguments to the conclusion that it is rationally permissible for me to take the left, unsweetened box.

Argument 3. I Know I have no Preference for the Contents of the Right Box

Being rational involves, at least in part, acting on preferences between outcomes. So, surely:

*Recognition:*

Whenever I have two options, and I know that I have no preference between the outcome of the one and the outcome of the other, it is rationally permissible for me to take either.

In this case, I know that I have no preference between the outcome of my taking the left box and the outcome of my taking the right box. Either there’s A in the left box, B+$1 in the right box – in which case I have no preference between the outcome of my taking the left box and the outcome of my taking the right box. Or there’s B in the left box, A+$1 in the right box – in which case I have no preference between the outcome of my taking the left box and the outcome of my taking the right box. So it is rationally permissible for me to take the left box.
Argument 4: It is Okay to Defer to My Better-Informed Self

Roughly: I know for sure that, if I were to see inside the boxes, I would have no preference for taking the right box. And it is rationally permissible for me to defer to my better-informed self.

More carefully: Thinking of a state of affairs as a way for things to be, and thinking of a maximal state of affairs as a precise way for everything to be, here are two very plausible principles concerning rational permissibility:

Deference

If I know that any fully informed, rational person, with all and only my preferences between maximal states of affairs, would have a certain array of preferences between sub-maximal states of affairs on my behalf, then it is rationally permissible for me to have that array of preferences between sub-maximal states of affairs.

Permissibility of Action Follows Permissibility of Preference

If I have just two options, and it is rationally permissible for me to have no preference for my taking the one, and no preference for my taking the other, then it is rationally permissible for me to take the one and rationally permissible for me to take the other.

In this case I know that any fully informed, rational person, with all and only my preferences between maximal states of affairs, would have no preference for my walking away with the right box. So, by Deference, it is rationally permissible for me to have no preference for walking away with the right box. So, by Permissibility of Action Follows
Permissibility of Preference, it is rationally permissible for me to walk away with the left box.

3.5 Take the Sweetened Option

We have two ways of thinking about rationality. Call the first (take the right box!) way of thinking prospectism and the second (take either box!) way of thinking deferentialism. Which of them is right? I think this is a difficult, open problem. I feel the pull of both. But, on balance, I lean towards prospectism. It is not that I have a dazzling, decisive argument that goes significantly beyond what I have said already. It is rather that I accept that the idea that being rational involves being sensitive to reasons, and I think that, by appealing to this idea we can explain why the deferentialist arguments are both attractive and wrong.

The deferentialist arguments are attractive because sometimes, when I learn that any fully-informed, rational person with all and only my preferences between maximal states of affairs (henceforth, for brevity: my better informed self) would lack an all things considered preference for my taking the right box, it does indeed cease to be true that I have a reason to take the right box and no reason to take the left box.

Suppose I learn that my better informed self would prefer that I take the left box. Now there remains something to be said for my taking the right box:

Pro-Right: If I take the right box then I will get that extra dollar.

And there is something new to be said for my taking the left box:

Pro-Left: My better informed self would prefer that I take the left box.
So the prospectist reasoning does not kick in. I am under no rational obligation to take the right box.

Or suppose I learn that my better-informed self would be all-things-considered indifferent in a sweetening *sensitive* way between my taking the left and right boxes (lightly sweeten either box and he would prefer it, lightly sour either box and he would dis-prefer it). Again there remains something to be said for my taking the right box:

Pro-Right: If I take the right box then I will get that extra dollar.

And there is something new to be said for my taking the left box:

Pro-Left: Call the collection of things in the left box that I do not know to be in the left box (which is to say everything in the left box) the *unknown contents of the left box*. Call the collection of things in the right box that I do not know to be in the right box (which is to say everything in the right box minus that dollar) the *unknown contents of the right box*. My better informed self would prefer to have the unknown contents of the left box to the unknown contents of the right box.

So, again, the prospectist reasoning does not kick in. I am under no rational obligation to take the right box.

The arguments are wrong because in the special case where I learn that my better-informed self would lack-all-considered preferences between my taking the left and right boxes in a sweetening *insensitive* way (lightly sweeten or sour either box and he would still lack a preference), the prospectist reasoning does kick in. In that case the something to be said for my taking the right box remains:
Pro-Right: If I take the right box then I will get that extra dollar.

But there is nothing new to be said for my taking the left box. I can’t say that my better informed self would prefer the contents of the left box to contents of the right box. I can’t say that he would prefer the invisible contents of the left box to the invisible contents of the left box. I have no reason to take the left box. So, in this special case, I have most reason to take the right box.

3.6 An Aside: Formal Prospectist Decision Theory

My goal in this chapter was just to make a case for the prospectist way of thinking about cases like Two Opaque Boxes. I am done with that. But in the last chapter I mentioned that classical expected utility theory is silent with respect to these questions because classical expected utility theory does not tell us what it is rationally permissible to do when preferences are negatively intransitive. You may be wondering, then, how we might extend classical expected utility theory so that it gives broadly prospectist answers about what it is rationally permissible to do when preferences are negatively intransitive (if you were not wondering this, and you have no patience for technical material, then please skip ahead to Chapter 4.)

This is fairly straight-forward. Here’s a rough statement of a decision theory I will call formal prospectism: We say that it is rationally permissible for me to take an action if and only if, for some way of rendering my preferences negatively transitive by keeping the preferences I have and adding new ones, the standard theory says that no alternative has higher expected utility.
Here’s a more accurate, formal statement of the theory. First some terminology:

Where \( U \) is a function that assigns numbers to outcomes, say that \( U \) represents a coherent completion of my preferences when, for all outcomes \( o_1, o_2 \), if I prefer \( o_1 \) to \( o_2 \), then \( U(o_1) > U(o_2) \), and for all prospects \( p_1, p_2 \), if I prefer \( p_1 \) to \( p_2 \) then \( \sum_o (P_{p_1}(o).U(o)) > \sum_o (P_{p_2}(o).U(o)) \), where \( o \) is a variable ranging over outcomes and ‘\( P_p(o) \)’ refers to the probability assigned to outcome \( o \) by prospect \( p \).

Now for the central claim:

**Formal Prospectism**

It is permissible for me to choose an option iff, for some utility function that represents a coherent completion of my preferences, \( U \), no alternative has greater expected \( U \)-utility.\(^{26} \)

In the Two Opaque Boxes case, formal prospectism says that it is rationally impermissible for me to take the left, unsweetened box. Why? Well, I prefer A+ to A, and B+ to B, so for any function, \( U \), that represents a coherent completion of my preferences, \( U(A+) > U(A) \), and \( U(B+) > U(B) \). So for any function, \( U \), that represents a coherent completion of my preferences, the expected \( U \)-utility of my taking the right box (\( 0.5U(B+) + 0.5U(A+) \)) is greater than the expected \( U \)-utility of my taking the left box (\( 0.5U(A) + 0.5U(B) \)).

---

\(^{26}\) What is the ‘expected \( U \)-utility’ of an act? You can interpret this in different ways, depending on how you feel about Newcomb problems, causal and evidential decision theory. If you wish to be a causalist prospectivist, then interpret it in a standard causalist way: the expected \( u \)-utility of act \( a \) is \( \sum_d (P(d).u(ad)) \), where \( d \) is a variable ranging over dependency hypotheses, propositions concerning how things beyond my control are, and ‘\( P(d) \)’ refers to my credence that hypothesis \( d \) is true, and ‘\( ad \)’ refers to the outcome of my taking act \( a \), if hypothesis \( d \) is true. If you wish to be an evidentialist prospectivist, then interpret it in a standard evidentialist way: \( \sum_d (P(d/a).u(ad)) \). This will not make a difference in any of the cases I discuss here.
3.7 A Yet-Further Aside: Formal Deferentialist Decision Theory

You may also be wondering about how we might extend classical expected utility theory so that it gives broadly deferentialist answers about what it is rationally permissible to do when preferences are negatively intransitive (again, if you are not wondering this, and you have no patience for technical material, please skip ahead to Chapter 4.)

The extension is much less straight-forward. A natural first move is to partition logical space into a set of dependency hypotheses – thinking of a dependency hypothesis as a maximally specific proposition concerning how things that matter to me causally depend on my present actions.\(^{27}\) As a notational matter, let ‘P(d)’ refer to my credence that dependency hypothesis d is true, and let ‘od’ refer to the outcome of my taking option o, if dependency hypothesis d is true. We can then restate the ‘Recognition’ principle from the last section in a more precise way:

**Recognition**

It is rationally permissible for me to choose option o if, for all alternatives open to me a, and all dependency hypotheses in which I have positive credence d, I do not prefer ad to od.

This is fine, so far as it goes, but it is not a general theory of decision, a theory that gives us full necessary and sufficient conditions for rational permissibility. It gives us one

---

\(^{27}\) This term was introduced by David Lewis in Lewis (1981). Other philosophers, and many decision theorists, talk of ‘states’ and ‘states of nature’.
sufficient condition for rational permissibility, but it tells us nothing about cases in which the condition does not apply. For example:

**Two More Opaque Boxes**

You show me items A and B, a coin, and two opaque boxes. Then you toss the coin and, governed by the toss, place item A in one box and item B in the other. I don’t see which item went where. Then, with some determinate probability, you either do or do not switch the item in the right box with item C — _where C is an item that I prefer to both A and B_. I don’t see whether you made the switch.

In this case I do not know that I have no preference for the outcome of my taking the right box. I do know that I have no preference for the outcome of my taking the _left box_.

But maybe the right box contains item C. If it does, then I have a preference for the outcome of my taking the right box. The Recognition principle is silent.

What do we want our general theory to say about this sort of case? I suggest that we want it to say the following:

**Mild Chancy Sweetening**

When I do not have a strong preference for C over A and B, and my credence that you made the switch is small, it is rationally permissible to take the left box. (The right box has been _mildly sweetened_ — not by a certain dollar, as in the original case, but by a small chance that it contains something that I regard as a little bit better than either A or B.)

Why? Well, we don’t want the theory to say that it is permissible to ignore a certain-dollar-sweetening, but impermissible to ignore (e.g.) a one-in-a-million-chance-of-a-
hundred-dollars-sweetening. I far prefer a certain dollar to a one in a million chance of a hundred dollars.

*Powerful Chancy Sweetening*

When I have a strong preference for C over A and B, and my credence that you made the switch is large, it is rationally impermissible to take the left box. (The right box has been *powerfully sweetened* – by a large chance that it contains something that I regard as much better than A or B.)

Why? Well, obviously, if I am almost certain that the right box contains C, and C is a billion dollars, then I ought to take it.

Is there a moderately natural, general theory of decision that says all these things? I think so. Here’s a rough statement of the theory I will call *formal deferentialism*: To find out if an action is permissible, I go to each relevant dependency hypothesis in turn. I take the coherent completion of my preferences that is most flattering to the action, supposing that the dependency hypothesis is true. I assign utilities to each of the actions open to me accordingly. I multiply these utilities by my credence that the dependency hypothesis is true… and move on to the next dependency hypothesis. I sum up. If there is some way of doing this on which the action comes out ahead of (or at least *not behind*) its competitors, then the action is permissible.

Here’s a more accurate, formal statement of the theory. First some terminology: Let C be the set of utility functions that represent coherent completions of my preferences. Let a *regimentation* of C be a subset, R, of C such that for some outcomes A, B, for any function $g$, $g \in R$ iff $g \in C$ and $g(A) = 1$ and $g(B) = 0$. (Note that it follows from the fact that
utility functions are unique under positive affine transformation\(^{28}\) that if \(R\) is a regimentation of \(C\), then for each coherent completion of my preferences, \(R\) has one\(^{29}\) and only one\(^{30}\) representative of it as a member.)

Some more terminology: For any regimentation \(R\), let the dependency-expansion of \(R\) be the set of functions \(f\) such that for any dependency hypotheses \(d\), for some function \(r\) in \(R\), for all actions open to me, \(a, f(ad) = r(ad)\). (Each function in the dependency-expansion of \(R\), for each dependency hypothesis, agrees with some function in \(R\) on the utilities of the states of affairs that will come about if I act one way or another and the dependency hypothesis is true.)

Now for the central claim:

*Formal Deferentialism*

It is permissible for me to choose an option iff, for some regimentation, \(R\), of the set of utility functions that represent my preferences, for some function \(r\) in the dependency-expansion of \(R\), no alternative has higher expected \(r\)-utility\(^{31}\).

In our original Two Opaque Boxes case, deferentialism says that it is rationally permissible to take the left, unsweetened box. In that case there are two relevant dependency hypotheses. According to the first, the left box contains A, the right B+.

---

\(^{28}\) Function \(v\) represents the same complete, coherent preferences as function \(u\) iff for some number \(i>0\), for some number \(j\), for all \(x\), \(v(x) = iu(x)+j\).

\(^{29}\) Proof: Take any coherent completion of my preferences, and a function that represents it, \(g\). Now let \(h\) be the function such that for all \(x\), \(h(x) = (1/(g(A)-g(B))) \cdot (g(x) - (g(B)/(g(A)-g(B))))\). By construction, \(h(A)=1\), \(h(B)=0\). And \(h\) represents the same coherent completion of my preferences as \(g\), because for some number \(i>0\), for some number \(j\), for all outcomes \(x\), \(g(x) = ih(x)+j\).

\(^{30}\) Proof: Suppose that \(g\) and \(h\) are functions in \(R\) that represent the same coherent completion of my preferences. Because \(g\) and \(h\) are functions in \(R\), \(g(A)=h(A)=1\), and \(g(B)=h(B)=0\). Because \(g\) and \(h\) represent the same coherent completion of my preferences, for some number \(i>0\), for some number \(j\), for all outcomes \(x\), \(g(x) = ih(x)+j\). Solving for \(i\) and \(j\), \(i=1\) and \(j=0\). Function \(g\) is function \(h\).

\(^{31}\) Again, please feel free to interpret ‘expected \(r\)-utility’ in the causalist or evidentialist way, depending on your feelings about Newcomb problems.
According to the second, the left box contains B, the right A+. For any regimentation, R, of the utility functions that represent coherent completions of my preferences, some functions in R assign A utility greater than or equal to B+, others assign B utility greater than or equal to A+. Choose one of the former, call it $U_1$, and one of the latter, call it $U_2$. Notice that $(0.5(U_1(A)) + 0.5(U_2(B))) \geq (0.5(U_1(B+)) + 0.5(U_2(A+)))$, so for some function $f$ in the dependency-expansion of R, $(0.5(f(A)) + 0.5(f(B))) \geq (0.5(f(B+)) + 0.5(f(A+)))$, so for some function $f$ in the dependency-expansion of R, the expected $f$-utility of taking the left box is greater than or equal to the expected $f$-utility of taking the right box.

And, more generally, deferentialism entails the principles I have called Recognition\textsuperscript{32}, Mild Chancy Sweetening, and Strong Chancy Sweetening.

\textsuperscript{32} To be accurate: If we plug the causalist expected-utility formula into Deferentialism then it entails the Recognition principle simpliciter. If we plug the evidentialist expected-utility formula into Deferentialism then it entails the Recognition principle except in Newcomb cases – cases where I care about how things out of my control are, and my conditional credence in things out of my control being one way or another, varies with actions available to me.
Chapter 5: The Same-Number Non-Identity Problem

5.1 A Canonic Presentation of the Problem

Remember Mary, the mother we talked about in the introduction:

Not Postponing Conception
Mary is recovering from German measles. Her doctor recommends that she postpone her efforts to conceive a child for a couple of months. If she conceives a child in the next couple of months then the child will, most likely, have moderate health problems. Mary has no strong reasons to conceive a child immediately, but she does have a mild preference for getting on with it. She gets on with it. Nine and a half months later baby Mariette is born, with significant health problems. This is not a disaster – Mary is a woman of means, so Mariette’s health problems do not impose a burden on wider society, and, on balance, Mariette has a rewarding life. But still, Mariette’s health problems are a chronic source of anxiety, pain and frustration to her.

Most people, when they think about Mary’s behavior, have a more or less inchoate sense that something has gone awry, morally speaking. To put it in more precise terms than may be warranted by the inchoate sense: Mary does something wrong. She ought to have followed her doctor’s advice, and waited a couple of months.

Why does Mary do something wrong? As I said earlier, normative ethicists, following Derek Parfit, would call this an instance of the ‘non-identity problem’, more specifically the ‘same-number non-identity problem’ (to distinguish it from the problem of what to say about cases in which our choices influence not just which people will exist, but also how many people will exist.) They think it deserves the name ‘problem’, rather than ‘question’ or ‘query’, because many of the obvious answers turn out to be
unsatisfactory. I will briefly survey these answers here, and some reasons why normative ethicists have taken them to be unsatisfactory.

First Answer: Mary does wrong by making things worse for Mariette. It would have been better for Mariette if Mary had waited.

The problem with this answer is that, as we saw in the Introduction, it seems plausible to think that Mariette would not have been better off if Mary had waited, because Mariette would not have existed if Mary had waited. More carefully, the following all seem true:

Reshuffling the Genetic Deck
Most probably, if Mary had waited two months to conceive a baby, then she would have conceived a baby as genetically different from the actual Mariette as typical non-twin siblings are different.

Genetic Essentialism
If Mary had conceived a baby as genetically different from the actual Mariette as typical non-twin siblings are different, then Mariette would never have existed.

Existence does not Harm Mariette
It would not have been better for Mariette if she had never existed.

It follows that, most probably, Mary has not made Mariette worse off. Most probably it would not have been better for Mariette if Mary had waited.

3 I should mention that at least one philosopher denies this. See Benatar (1997), an essay aptly titled: ‘Why it is Better Never to Come Into Existence’. This is an almost comically bleak view.
Second Answer: Mary does wrong by making things worse. It would have been better (not better for Mariette, just better) if Mary had waited.

This was Parfit’s solution to his problem. He appealed to a principle he called ‘The Same Number Quality Claim, or Q’:

If, in either of two possible outcomes the same number of people would ever live, it would be worse if those who live are worse off, or have a lower quality of life, than those who have ever lived.\(^{34}\)

If Mary had waited then she would have conceived a child who would have been better off than Mariette actually is. So, by Q, things would have been better if Mary had waited. So she did wrong by failing to wait.

There are at least two reasons why many philosophers find Parfit’s answer prima facie unsatisfactory. The first is that is far from obvious that Mary has brought about a worse outcome by failing to wait. Obviously those philosophers who were moved by the ‘separateness of persons’ objection to utilitarianism will not accept this. They will say that when we have two outcomes, one better for one person, the other better for another person, there is no interesting sense in which one can be just better than another. But the objectors need not be so radical. A quieter, and I think more sensible, objection is that Mary does not bring about a worse outcome in virtue of incommensurability in the values that determine who is better off than who. Consider three outcomes:

M—: (the actual outcome) in which Mary conceives Mariette and Mariette has an on-balance-good-life, mildly blighted by health problems.

---

\(^{34}\) Parfit (1984), 360.
M: in which Mary conceives Mariette and Mariette has a life just like her actual life, but without the health problems.

J: in which Mary conceives John, a child very different from Mary, with different passions, habits, patterns of thinking, and John goes on to live a healthy life.

It may be that Mariette in M is better off than Mariette in M—. It may also be that Mariette in M is no better or worse off than John in J. But it does not follow that John in J is better off than Mariette in M—. Just as there is incommensurability in the values that determine which of two restaurants is better, so there is incommensurability in the values that determine which of two people is better off.

Most probably, if Mary had waited to conceive a child, then she would have brought about an outcome like J. So, most probably, she would have brought about an outcome no better or worse than the actual outcome.

The second reason to find Parfit’s answer prima facie unsatisfactory is that it is far from obvious that, when we make decisions about procreation, we have a standing obligation to make the world a better place. My wife and I did not conceive a child last year. Suppose that the world would have been better if we had conceived a child, not in virtue of the benefits that the child would have bestowed on actual people, but in virtue of the fact that the child would have been happy and healthy and fun-loving and so forth. Suppose that we were in a position to know that. Does it follow that we did something wrong? I say it does not! This is a sparkingly clear example of a case in which traditional act consequentialism seems too demanding. We are under no obligation to make the world better by creating happy children if we do not feel like doing so. So why does
Mary have an obligation to make the world better by creating happier children if she does not feel like doing so?  

Moved by worries like this, several philosophers have returned to the idea that Mary’s wrong is somehow personal. It isn’t about how she stands in relation to the world. It’s about how she stands in relation to Mariette.

*Third Answer:* Mary does wrong by somehow wronging Mariette, though she does not make Mariette worse off.

There are many ways to develop this answer. One is to say that Mary has wronged Mariette by violating Mariette’s rights — you can violate somebody’s rights without making them on balance worse off. Another is to say that Mary has harmed Mariette in an illegitimate way — you can harm someone in an illegitimate way without making them worse off.

One concern about these proposals is that they seem to invoke non-standard notions of ‘right’ and ‘harm’. What right of Mariette’s does Mary violate? A right-not-to-be-born-with-health-problems-when-the-alternative-is-never-being-born-at-all? Do we

---

35 I should note that Parfit is agnostic about whether we make the world better by creating happy children. He showed, in the last part of Parfit (1983), that some formidable obstacles stand in the way of a full theory of when worlds containing different numbers of people are better or worse than one another. But that does not block the point being made here. The point is conditional: Suppose I would have made the world better by conceiving a child last year. It does not follow that I did wrong by holding back.

36 This broad approach was taken by James Woodward in Woodward (1986), and, more recently, by David Velleman in Velleman (1986).

37 Variants of this approach are taken by Seana Schiffrin (see Schiffrin 1999), by B. Steinbock and R. McClamrock, (see Steinbock and Mclamrock 1994), and by Elizabeth Harman (see Harman 2004).

38 David Velleman talks of a ‘right to be born into good enough circumstances’ (Velleman 2008: 275). He also says “A child to who we give a lesser initial provision will have been wronged by our lack of due concern for human life in creating him – our lack of concern for human life itself, albeit in his case.’ (ibid.: 276) I don’t think it necessary to show that Mary has violated any right of Mariette’s, but, as you will see in the next section, I agree that it is helpful to focus on the attitudes of the parents.
really have such rights? And can we really harm someone when they are predictably much better off as result of what we do? Can this, the fact that we ‘harmed’ them in this odd way, really explain why we do wrong?

Another concern is that the proposals struggle to distinguish between cases that really should be distinguished. Contrast Not Postponing Conception with this case:

**Going Ahead**

Mary is trying to conceive a child. Her doctor tells her that if she succeeds then her child will have moderate health problems – “You can’t avoid that. It’s genetics.” Mary doesn’t let this cramp her style. Nine and a half months later baby Mariette is born, with just the health problems that Mary’s doctor foresaw. This is not a disaster – on balance, Mariette has a rewarding life. But Mariette’s health problems are a chronic source of anxiety, pain and frustration to her.

If Mary harms Mariette, or violates her rights, in Not Postponing Conception, then it would seem that she does the same in Going Ahead. After all, the only difference between the cases is that in Not Postponing Conception there is a salient counterfactual alternative in which Mary conceives a different, healthy child – and that difference does not seem to bear on whether Mary harms Mariette or violates her rights. But we don’t want to say that she does wrong or wrongs Mariette in Going Ahead.39

---

39 Parfit made a version of this point in response to Woodward in Parfit (1986). Schiffrin and Harman are both aware of it. Schiffrin appears to embrace the conclusion that Mary does wrong in the second case, and run with it – the harms that we bring upon our progeny by creating them are not, in general, balanced by the benefits of existence. Harman argues that, to explain why Mary does wrong in the first case but not in the second, we need to focus not just on the harm she did, but also on her reasons for action. In the first case she has weak reasons not to wait. In the second case she has strong reasons to go ahead (she wants a baby, we may assume, and this is the only way to have one). As you will see in the next section, I don’t think it necessary to show that Mary has harmed Mariette in these cases, but I agree that it is helpful to focus on Mary’s reasons for action.
Will no straight answer work? Some philosophers have come to think so. They say that, after thirty years of failing to answer the question, we should conclude that it contains a false presupposition.

*Fourth Answer:* In fact Mary does no wrong. Once again, careful reflection shows us that a powerful moral intuition is misguided.\(^{40}\)

This is a conclusion-of-the-last-resort. Let’s see if we can avoid it.

### 5.2 Solving the Same-Number Non-Identity Problem

Happily, the assumptions we have made about benevolence and rationality give us an easy solution to the same-number non-identity problem. As a warm-up, consider a simpler, very unrealistic version of Mary’s case.

#### Not Postponing Conception (simplified)

As before, but this time Mary takes herself to be capable of conceiving only two children – Mariette and Jimmy. Mary regards it as equally likely that she will conceive either, whether or not she goes ahead or waits.

In this case, supposing that decency requires of Mary that she be minimally benevolent towards her future children and rationality requires of her that she follow the recommendations of prospectism, if Mary is decent and rational then she will wait.

---

\(^{40}\) This view is taken by Melinda Roberts in Roberts (1998), and by David Boonin in Boonin (2008). Both philosophers work hard to persuade us that our intuitions about cases like Mary’s are distorted by nearby cases, cases in which Mary does make somebody worse off.
The laborious way to see this is to observe that Mary has two relevantly different options:

- Going full-steam ahead
- Waiting

And there are four relevantly different things that might end up happening:

- **M—**: in which Mary conceives Mariette, and she is unhealthy.
- **J—**: in which Mary conceives Jimmy, and he is unhealthy.
- **M**: in which Mary conceives Mariette, and she is healthy.
- **J**: in which Mary conceives Jimmy, and he is healthy.

The prospect associated with going full-steam ahead is \( \frac{1}{2} \) each of M— and J—, while the prospect associated with waiting is \( \frac{1}{2} \) each of M and J. Supposing Mary is minimally benevolent towards the children she is in a position to have, she must prefer M to M—, J to J—. So, for any completion of her preferences, waiting must have higher expected utility than going full steam ahead. So, by prospectism, it is rationally impermissible for her to go ahead.

The quick way to see it is to note that the prospects associated with the options open to her in this case are just the same as the prospects associated with the options open to her in:

**Not Taking the Pill (Simplified)**

Mary is pregnant, and recovering from German measles. Her doctor recommends that she take some pills. If she does not then her child will, most likely, have significant health problems. If she does then her child will, most likely, be healthy. She takes herself to be have been capable of conceiving
only two children – Mariette and Jimmy. She regards it as equally likely that either child is in her belly, whether or not she takes the pill.

And obviously Mary will take the pill in this case, if she is minimally benevolent towards her children and rational.

Now, in the realistic case the prospects associated with Mary’s options are much more complicated (she has little idea what child she will conceive if she goes full-steam ahead, little idea what child she will conceive if she waits). But, just so long as the following as true:

Invariance For any given child she might have, she regards it as no more or less likely that she will have that child if she waits than if she goes full-steam ahead.

then she will wait, if she is minimally benevolent and rational. The quick way to see why this is so, is to note that, if Invariance is true, the prospects associated with the options available to her are relevantly just the same as the prospects associated with the options available to her in another case:

Not Taking the Pill Again she is pregnant, and recovering from German measles. Again her doctor recommends that she take some pills. If she does not then her child will, most likely, have significant health problems. If she does then her child will, most likely, be healthy. But this time (as with real pregnancies) she does not have restricted ideas about who she might have conceived.
And, obviously she will take the pill in this case, if she is both minimally benevolent towards her children and rational.

5.3 Mariette’s Complaint

So that’s the solution to the same-number non-identity problem: if Mary were decent and rational then she would have waited. Given the way she behaved, there must have something defective about her – she was either rationally defective or morally defective.

Does it imply that Mary has somehow wronged Mariette? Well, a nice feature of the solution is its explanation of why Mariette, of all the people in the world, has a special complaint against Mary. Mariette cannot say: “You made me worse off.” But she can say this: “Supposing that you acted rationally, you must not have been minimally benevolent towards all the children you were in a position to conceive. In particular (supposing you were not positively ill-willed towards any of the children you were in a position to conceive) you must not have been minimally benevolent towards me. You must not have preferred state of affairs $M$, in which I am healthy and conceived three months later, to state of affairs $M^-$, in which I am unhealthy and conceived three months earlier. That’s not just bad parenting. It’s an improper indifference to me, an attitude that I am uniquely qualified to resent and complain about.”

5.4 Objection: Again, What About Fairness?

“But wait” you might say. “What about Taurek and fairness? Doesn’t fairness demand that we flip a coin when lives are at stake? Indeed, whether or not Taurek is right
about saving one or many, isn’t it manifestly unfair to let mild health issues be a tie-breaker when lives are at stake? Consider one more rescue case:

**Disease-Based Small Differences, Opaque Inter-Personal Conflict**

As before, I know that one person is on an island to my west, another person on an island to my east, and I know almost nothing about these people. But this time both islands are three hours away. The only asymmetry I am aware of is this: The person on the western island has had health problems since birth.

Suppose I head east, on the grounds that the person on the eastern island will most probably be healthier, if I save him or her, then the person on the western island will be, if I save him or her. Then the western-person might complain: “This is desperately unfair! First I had to live with health problems. Then you used this very fact as grounds for saving eastern-person and leaving me to die. It is unfair to use my misfortune as grounds for further disadvantaging me.”

Has the western-person got a point? Maybe so. But even if we concede the point, there is an important difference between this case and the Not-Taking the Pill case. In the Not-Taking the Pill case, if Mary waits to conceive her child, there is nobody unfairly discriminated against on grounds of his or her health, because the child Mary would have conceived if she had gone full-steam ahead never exists. These issues of fairness just don’t arise when we think about procreation.
5.5 Objection: Contrary Desires

“But wait” you might say. “The argument only applies when Mary takes herself to have *no* reason to go full-steam ahead. Just so long as there is something else at stake for her, no matter how fickle (Scorpio-babies are just *so hip* right now!), it does not follow from what we have said that if she is benevolent and rational then she will heed her doctor’s advice.”

True, if we are to show that decency and rationality commit her to waiting in this new case, then we will need to assume that, to be decent, you must be more than just minimally considerate of, and minimally benevolent towards the children you are in a position to have. You must also take considerations to do with the interests of the children you are in a position to have significant weight when they conflict with other considerations.

This is a further assumption, so it deserves another **FLAG**. But it is a safe assumption. Consider:

**Not Taking the Mildly Unpleasant Pill**

Again Mary is pregnant, and recovering from German measles. Her doctor recommends that she take some pills. If she does not then her as-yet-unborn child will, most likely, have significant health problems. If she does then her child will, most likely, be healthy. But this time the taking of the pills is very mildly unpleasant to her. They won’t make her sick, but they are large, and swallowing them is a bit of a chore.

We all think that in this case, if Mary is decent then she will take the consideration ‘taking the pills will be good for my as-yet-unborn child’ to be decisive. Her all-things-
considered preference will be for chore-somely taking the pills and giving birth to a healthy child.

If this assumption is correct, if decency requires of you that you take considerations to do with the interests towards your children to have significant weight, then, by the same arguments as before, decency and rationality together require of Mary that she heed her doctor’s advice when doing so will frustrate her mild desire to have a Scorpio-baby.

Just how far does this go? Just how considerate and benevolent must we be? Do decency and rationality together require of Mary that she heed her doctor’s advice in these cases?

Not Postponing Conception for a Long Time

Once again Mary’s doctor recommends that she postpone her efforts to conceive a child, otherwise her child will most likely have significant health problems. But this time it’s a two-year postponement, and she is at the tail end of her fertility curve. If she waits then she may never have a child at all.

Not Taking the Sickening Pill

Once again Mary is pregnant, and her taking some pills will, most likely, save her child from significant health problems. But this time there is a significant chance that the pills will damage Mary’s liver.

I don’t know. We have reached a point where platitudes about benevolence and practical rationality will not help us.
5.5 Objection: Desires About People Need People

“But wait” someone might say. “We have assumed that decency requires of Mary that, prior to conception, she be benevolent towards her future children, that she prefer, for example, that Jimmy live a better life than a worse life. But Jimmy does not, and will never, exist! You can’t have desires about things that never exist.”

Maybe this is right. If so, then it is not right to say that decency requires of Mary that she have desires about Jimmy. But we can still say that it requires of her that she have some desires in the close vicinity. Consider my son Inigo. He is, touch wood, quite healthy. It might not have been like this. He might have come down with Polio last year. Contrast a richly detailed description of the life he would have led if he had caught Polio last year with a richly detailed description of his actual life. Suppose these descriptions had been presented to me before Inigo existed. ‘Here are two ways things might go: First, you conceive a child on this date… genetically like so… and you name him ‘Inigo’… and he comes down with Polio, suffering terribly… Second, you conceive a child on the same date… genetically just the same… and you name him ‘Inigo’… and he does not come down with Polio...’ I think, that though I may have been incapable of forming desires about Inigo, in particular, I would and should, on pain of indecency, have desired that things go the second way. And that’s all we need.

5.6 Objection: Essentialism About Origins

“But wait” someone might say. “If Mary understands what makes people who they are, then Invariance will not be true of her in the non-identity case. That principle says:
Invariance  For any given child Mary might have, she regards it as no more or less likely that she will have that child if she waits than if she plunges full-steam ahead.

But people are essentially around the times at which they are conceived – they could not have been conceived at other times. Take Elizabeth II, for example. She was conceived in July 1925. If her parents had conceived a baby six months earlier then, even if the baby had been genetically just the way she actually is, even if the baby had gone on to have a life just like her actual life (the crown, the solid pastel overcoats, the dour expression and so forth), Elizabeth would never have existed. If Mary understands this deep fact about our essences then she will not regard it as no more or less likely, for any given child she might have, that she will conceive that child if she waits than if she goes full steam ahead. She will think, rather, like this. “Look, there’s a range of different children I might conceive if I go full-steam ahead. Choose one of them, call him John, with a particular genetic profile, call it G-John. There’s a very small chance that I will conceive John if I go full steam ahead, but there’s absolutely no chance that I will conceive John if I wait, because John can only be conceived this month. True, if I wait then there’s a very small, equal chance that I will have a child with the genetic profile G-John, but this child would not be John. He would be a qualitatively similar, numerically distinct child.”

This is a funny view about essence and possible children. I think that, if my parents had conceived a child in January 1971, genetically just like me, and that child had gone on to live a life just like mine, up to the point where he wrote a book just like this one, then I would have been conceived six months before I was actually conceived. Of course,
it is very unlikely that this would have happened, if they had conceived a child in January 1971. But if they had somehow managed it then the baby would have been me.

No matter. I will show, in the second part of this book, that even if Mary adopts the highly unobvious view about essence and possible children, if she is decent and rational then she will heed your doctor’s advice in the non-identity case. This is because, if she is decent and rational then she will prefer that she conceive a particular healthy child at a certain time, rather than conceive a genetically identical, numerically distinct, unhealthy child some time earlier.

My argument will also allow us to cut yet further into the jungle of normative ethics, to address cases in which I know precisely who I am in a position to benefit by doing what.