# ON THE OVERWHELMING IMPORTANCE OF SHAPING THE FAR FUTURE

By

NICHOLAS BECKSTEAD

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Philosophy

Written under the direction of

Larry Temkin

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2013

# ABSTRACT OF THE DISSERTATION

## On the Overwhelming Importance of Shaping the Far Future

by NICHOLAS BECKSTEAD

Dissertation Director:

Larry Temkin

In slogan form, the thesis of this dissertation is that shaping the far future is overwhelmingly important. More precisely, I argue that:

> *Main Thesis*: From a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

The first chapter introduces some key concepts, clarifies the main thesis, and outlines what follows in later chapters. Some of the key concepts include: existential risk, the world's development trajectory, proximate benefits and ripple effects, speeding up development, trajectory changes, and the distinction between broad and targeted attempts to shape the far future. The second chapter is a defense of some methodological assumptions for developing normative theories which makes my thesis more plausible. In the third chapter, I introduce and begin to defend some key empirical and normative assumptions which, if true, strongly support my main thesis. In the fourth and fifth chapters, I argue against two of the strongest objections to my arguments. These objections come from population ethics, and are based on Person-Affecting Views and views according to which additional lives have diminishing marginal value. I argue that these views face extreme difficulties

and cannot plausibly be used to rebut my arguments. In the sixth and seventh chapters, I discuss a decision-theoretic paradox which is relevant to my arguments. The simplest plausible theoretical assumptions which support my main thesis imply a view I call *fanaticism*, according to which any non-zero probability of an infinitely good outcome, no matter how small, is better than any probability of a finitely good outcome. I argue that denying fanaticism is inconsistent with other normative principles that seem very obvious, so that we are faced with a paradox. I have no solution to the paradox; I instead argue that we should continue to use our inconsistent principles, but we should use them tastefully. We should do this because, currently, we know of no consistent set of principles which does better.

# Acknowledgements

I have received helpful comments from Geoff Anders, Gustaf Arrhenius, Nir Eyal, Katja Grace, Preston Greene, Beth Henzel, Doug Husak, Mark Lee, Ben Levinstein, Holly Morgan, Toby Ord, Andrew Sepielli, Jonah Sinick, and Holly Smith. I received extensive and helpful written feedback from Tim Campbell, Johann Frick, Theron Pummer, Carl Shulman, Gerard Vong, Matt Wage, and Alec Walen.

In the early stages of writing this dissertation, I got extensive feedback on a few key chapters, and my ideas for the project as a whole, from John Broome and Nick Bostrom. Will McAskill (formerly Will Crouch) read every single chapter and gave me great feedback.

I am also grateful to my excellent committee. I'm especially indebted to Ruth Chang, who helped me decide to write a dissertation in ethics. Derek Parfit and Jeff McMahan gave me very helpful feedback and brought some important considerations to my attention which would have been missed otherwise.

Larry Temkin has been an outstanding mentor and adviser over the last few years. His detailed and thoughtful comments on my work, both in written form and in the many afternoons and evenings we spent discussing my project, were immensely valuable. In addition to greatly improving this dissertation, working with Larry was excellent for my development as a writer and a thinker. He helped me become more modest and careful about my claims, and clearer in my own mind about what my arguments could and couldn't show. For that and more, I am very grateful.

I'm most grateful to Mikaela Provost, who has helped make the years spent writing this dissertation some of the happiest of my life.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction and Overview of the Dissertation

## 1.1 What is the question?

### 1.1.1 The rough future-shaping argument

How important is what happens in the far future? And how important is it to try to make sure that the far future goes well, in comparison with more ordinary concerns, such as improving lives of currently existing people? These questions often arise in the context of measuring the cost of climate change to future generations, and, in a way that is less well-known, they arise in thinking about the importance of preventing catastrophes that could cause human extinction or otherwise shape the far future.

In slogan form, the thesis of this dissertation is that shaping the far future is overwhelmingly important. More precisely, I mean to argue that:

> *Main Thesis*: From a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

A rough outline of my reasoning—where each of the steps require more elaboration and defense—is as follows:

1. Humanity may survive for millions, billions, or trillions of years.

2. If humanity may survive may survive for millions, billions, or trillions of years, then the expected value of the future is astronomically great.

3. Some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways.

4. If the expected value of the future is astronomically great and some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways, then from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

5. Therefore, from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

Several clarifications are in order. First, here and elsewhere in this dissertation, when I use words like "important," "good," "better," "best," and "matters," I am only talking about good and bad consequences for the world in general. I am not talking about other moral considerations. Second, by "humanity" and "our descendants" I don't just mean the species *homo sapiens*. I mean to include any valuable successors we might have. Third, the argument depends on some important empirical elaboration, and it's not something you could expect to figure out without a lot of empirical information. Fourth, when I say that shaping the far future is overwhelmingly important, I don't mean to suggest that it would be helpful if everyone were always thinking about what would be best for the far future. This may be analogous to the way in which our lives go better if we focus primarily on external aims than if we focus primarily on making our lives go as well as possible, or the way in which a financial trader might make better returns by operating with a simple set of heuristics than by trying to calculate expected returns on every transaction. Fifth, by "in expectation," I mean to highlight that I am talking about goodness of prospects relative to our uncertainty, and that expected values are highly relevant. Sixth, when I say "from a global perspective," I mean to be talking about general priorities for the world in general. What it would be best for a particular individual to do could vary significantly more, so the general claim is harder to justify, though it may be true that, for many individuals, what matters most about their lives is how their actions are expected to shape the far future. Some parts of this argument are hard to make more precise just yet. I'll say more about what I mean by "the general trajectory" very soon, and more about

what "not-ridiculously-small" means in chapter 3. Suffice it to say that making the future go one millionth better, in expectation, would be large enough to count as "not-ridiculously-small."

How convinced should you be by the arguments I'm going to give? I'm defending an unconventional thesis and my support for that thesis comes from highly speculative arguments. I don't have great confidence in my thesis, or claim that others should. But I am convinced that it *could well be true*, that the vast majority of thoughtful people give the claim less credence that they should, and that it is worth thinking about more carefully. I aim to make the reader justified in taking a similar attitude.

### 1.1.2 How could we affect the far future?

Thinking about how we could affect the far future helps clarify what I mean by "shaping the far future," illustrates some possible ways we could shape the far future in "not-ridiculously-small" ways, and helps relate my work to the ideas of other philosophers.

#### 1.1.2.1 Ripple effects of ordinary actions

Suppose I cure some child's blindness. We ordinarily think that the main benefit of this is that the child will have a better life. We may acknowledge that the child may be more economically successful, his parents will have more free time, and his school will spend less resources educating him. We may also acknowledge the benefits that other people enjoy because the child will be a more productive worker, his parents will have time for other activities, and it will take less resources to educate him. But typically, we do not pay much attention to these indirect effects, and pay very little attention to how they might propagate further into the future. However, curing the child's blindness creates a ripple effect that carries forward through many people affected by the child, and many people those people affect, and so forth. If these ripple effects continue for a long time, it can be argued that these they swamp the proximate benefits that command our attention.

This idea is implicit in the practice of instrumental economic discounting—a practice which is much more enlightened than non-instrumental discounting, despite various imperfections. Since the Industrial Revolution, many economies have been growing at exponential rates, a plausible mechanism for this is chains of human empowerment sparked by innovation, trade, moral development, and potentially many other unknown causes. But somehow or other, many countries have had steady growth over the last 200 years. And it is believable that if these economies were larger in the beginning, the additional resources would have compounded at a comparable rate. Instrumental

economic discounting assumes that economic benefits typically ripple forward like this, growing at an exponential rate, and it implies that the main benefit of curing a child's blindness will be the aggregate effects on future generations.

We may or may not be mistaken in focusing only on the proximate benefits. As in our financial trader example, it may often be that "maximize predictable benefits to people alive today" is better advice than "maximize benefits to future generations," even if the second piece of advice has a sounder theoretical basis. And it may be that altruists accomplish their goals no worse if they follow the second piece of advice rather than the first. In some cases, however, the two goals come apart. To take a couple of toy examples, hospice care and anesthetics for animals may have predictable short run benefits, but these benefits arguably have much less significant ripple effects than, say, improving people's computers.

#### 1.1.2.2 Existential risk reduction

Several potential catastrophes have a chance of destroying human civilization in the next century. Most of these threats—which include catastrophic climate change, future versions of bioterrorism, nuclear war, and some scenarios involving dangerous future technologies—involve a small, hard-to-estimate chance of our collective destruction, and a lack of clarity about the costs of reducing the risk (Posner, 2004).[1] Though expert assessments inherently involve subjectivity and selection bias, some people who have studied the issues carefully put the odds of such a catastrophe during the next century above 10% (Sandberg and Bostrom, 2008; Bostrom, 2012). These threats are examples of what Bostrom (2002) calls *existential risks*. Bostrom defines an existential risk as a risk "that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development," (Bostrom, 2012). If one of these threats destroys the human race or drastically limits its future potential, it would be an *existential catastrophe*.

"Do what is best for future generations" and "maximize predictable benefits to people alive today" may come apart very significantly in the case of existential risks. Bostrom, for example, accepts a version of the rough future-shaping argument, but adds the further empirical claim that:

> "[T]he loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole. It may be useful to

[1] See Bostrom and Ćirković (2008) for a survey of the empirical information relevant to many of the most likely scenarios.

adopt the following rule of thumb for such impersonal moral action:

*Maxipok:* Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe." (Bostrom, 2012, p. 10).

In other words, Bostrom argues that shaping the far future is overwhelmingly important, and that the best way of shaping the far future is to try to minimize existential risk.

Derek Parfit has also stressed the importance of humanity's future potential in the context of existential risk. He closed *Reasons and Persons* by arguing that the worst aspect of human extinction from a nuclear war would *not* be the losses to all the people alive at the time. Instead, he claimed, the greatest loss would be the opportunity cost of cutting humanity short—a loss of trillions or more potential lives, and all of humanity's future achievements (Parfit, 1984, pp. 453-454). And he closed the second volume of *On What Matters* with some reflections about the potential value of the future. I've reproduced some of Parfit's more striking thoughts below:

> We live during the hinge of history. Given the scientific and technological discoveries of the last two centuries, the world has never changed as fast. We shall soon have even greater powers to transform, not only our surroundings, but ourselves and our successors. If we act wisely in the next few centuries, humanity will survive its most dangerous and decisive period. Our descendants could, if necessary, go elsewhere, spreading through this galaxy. (Parfit, 2011, p. 616)

> The Earth may remain inhabitable for at least a billion years. What has occurred so far is at most a tiny fraction of possible human history. Nor should we restrict this question to the lives of future human beings. Just as we had ancestors who were not human, we may have descendants who will not be human... Our descendants might, I believe, make the further future very good. (Parfit, 2011, p. 616)

> What now matters most is that we avoid ending human history. If there are no rational beings elsewhere, it may depend on us and our successors whether it will all be worth it. (Parfit, 2011, p. 620)

Parfit does not go as far as Bostrom, but seems to believe that reducing existential risk is likely to be one of the most important issues in the world.

In my view, trying to reduce existential risk is potentially the most attractive way of trying to shape the far future because it works even if we know very little about what the far future will be like; there are genuine threats; there are not great market, governmental, and philanthropic incen-

tives to address the threats; and addressing small, hypothetical existential risks may be relatively unimportant by ordinary standards. It is also an attractive option to talk about because it is easy to understand in comparison with other ways of trying to affect the far future. Because of this, much of this dissertation will talk about how important it is to reduce existential risk.

#### 1.1.2.3 Trajectory changes

What really makes the arguments for the importance of existential risk reduction tick is (i) the fact that the future could be overwhelmingly good, and (ii) the fact that an existential risk of $x\%$ means that the future has $x\%$ less expected value than it would if there were no risk. But reducing existential risk is not the only way to make the whole future better by some fraction. Teaching the next generation slightly better moral values, for instance, might have a similar effect, and not just because teaching the next generation slightly better moral values would result in decreased existential risk. It could create a small, persistent change in the far future along various dimensions, including private social norms, political systems, or other traditions.

In thinking about how we might affect the far future, I've found it useful to use the concept of the world's *development trajectory*, or just *trajectory* for short. The world's development trajectory, as I use the term, is a rough summary way the future will unfold over time. The summary includes various facts about the world that matter from a macro perspective, such as how rich people are, what technologies are available, how happy people are, how developed our science and culture is along various dimensions, and how well things are going all-things-considered at different points of time. It may help to think of trajectory as a collection of graphs, where each graph in the collection has time on the x-axis and one of these other variables on the y-axis.[2]

With that concept in place, consider three different types of benefits from doing good. First, doing something good might have *proximate* benefits—this the name I give to the fairly short-run, fairly predictable benefits that we think about when we cure some child's blindness, save a life, or help an old lady cross the street. Second, there are benefits from *speeding up development*. In many cases, ripple effects from good ordinary actions result in speeding up development in the sense that they make the world move along its trajectory more quickly. Saving some child's life might cause his country's economy to develop very slightly more quickly, or make certain technological or cultural innovations arrive more quickly. Third, in other cases, our actions may slightly or significantly alter

[2]If the future does not evolve deterministically enough, there may be many potential future trajectories, so talking about "the" trajectory may be somewhat misleading. This difficulty could be avoided if I changed each occurrence of "trajectory" to "probability distribution over possible trajectories," but that would not be sufficiently more enlightening to justify the repeated use of a cumbersome expression.

the world's development trajectory. I call these shifts *trajectory changes*. If we ever prevent an existential catastrophe, that would be an extreme example of a trajectory change. There may also be smaller trajectory changes. For example, if some species of dolphins that we really loved were destroyed, that would be a much smaller trajectory change.

The concept of a trajectory change is related to the concept of *path dependence* in the social sciences, though when I talk about trajectory changes I am interested in effects that persist much longer than standard examples of path dependence. A classic example of path dependence is our use of QWERTY keyboards. Our keyboards could have been arranged in any number of other possible ways. A large part of the explanation of why we use QWERTY keyboards is that it happened to be convenient for making typewriters, that a lot of people learned to use these keyboards, and there are advantages to having most people use the same kind of keyboard. In essence, there is path dependence whenever some aspect the world could easily have been arranged in way $X$, but it is arranged in way $Y$ due to something that happened in the past, and now it would be hard or impossible to switch to way $X$. Path dependence is especially interesting when way $X$ would have been better than way $Y$. Some political scientists have argued that path dependence is very common in politics. For example, in an influential paper (with over 3000 citations) Pierson (2000, p. 251) argues that:

> Specific patterns of timing and sequence matter; a wide range of social outcomes may be possible; large consequences may result from relatively small or contingent events; particular courses of action, once introduced, can be almost impossible to reverse; and consequently, political development is punctuated by critical moments or junctures that shape the basic contours of social life.

The concept of a trajectory change is also closely related to the concept of a *historical contingency*. If Thomas Edison had not invented the light bulb, someone else would have done it later. In this sense, it is not historically contingent that we have light bulbs, and the most obvious benefits from Thomas Edison inventing the light bulb are proximate benefits and benefits from speeding up development. Something analogous is probably true of many other technological innovations: computers, candles, wheelbarrows, object-oriented programming, and the printing press, to give an arbitrary list of examples. There have been other events that were historically contingent, and changed the course of history significantly. Potential examples include: the rise of Christianity, the creation of the US Constitution, and the influence of Marxism. Various aspects of Christian morality influence the world today in significant ways, but the fact that those aspects of morality, in exactly

those ways, were part of a dominant world religion was historically contingent. And therefore events like Jesus's death and Paul writing his epistles are examples of trajectory changes. Likewise, the US Constitution was the product of deliberation among a specific set of men, the document affects government policy today and will affect it for the foreseeable future, but it could easily have been a different document. And now that the document exists in its specific legal and historical context, it is challenging to make changes to it, so the change is somewhat self-reinforcing.

Very persistent trajectory changes that are not existential catastrophes, could have great significance for shaping the far future. Though it seems unlikely that the far future will inherit many of our institutions exactly as they are, it is not hard to believe that various aspects of the far future—including social norms, values, political systems, and technologies—will be path dependent on what happens now, and often in a suboptimal way. In general, it is reasonable to assume that if there is some problem that might exist in the future and we can do something to fix it now, future people would also be able to solve that problem. But if values or social norms change, they might not agree that some things we think are problems really are problems. Or, if a certain standards or conventions get sufficiently entrenched, some problems may be too expensive to be worth fixing.

Though thinking about these smaller trajectory changes may be as important as thinking about existential risk, the best ways to address smaller trajectory changes may be very different. For example, it may be reasonable to try to assess, in detail, questions like, "What are the largest specific existential risks?" or, "What are the most effective ways of reducing those specific risks?" In contrast, I would not find it as effective to try to make specific guesses about how we might create smaller positive trajectory changes because there are so many possibilities and many trajectory changes do not have significance that is predictable in advance. No one could have predicted the persistent ripple effects that Jesus's life had, for example. In other cases—such as the framing of the US Constitution—it's clear that a decision has trajectory change potential, but it would be hard to specify, in advance, which concrete measures should be taken. Because of this, promising ways to create positive trajectory changes in the world may be highly indirect. Improving education, improving our children's moral upbringing, improving science, improving our political system, spreading humanitarian values, or otherwise improving our collective wisdom as stewards of the future could create many small, unpredictable positive trajectory changes.

### 1.1.3 Philosophical questions I don't try to answer

I'm only asking how *good* it is to shape the far future, in an impartial sense. I do not address deontological considerations—such as whose responsibility that is, how much the current generation should be required to sacrifice for the sake of future generations, how shaping the far future stacks up against special obligations or issues of justice, or how people should make trade-offs between pursuing causes they are personally passionate about vs. shaping the far future (supposing that shaping the far future is, as I argue, overwhelmingly important). These are all good questions, and if I had quick answers to them, I'd offer them. But I don't have quick answers, and the questions I'm addressing here are hard enough on their own. Therefore, my focus is on answering the theoretical questions that need answering if we are to know how good it would be to shape the far future, in comparison with our best alternatives for doing good.

This does not mean that my arguments are only relevant to consequentialists. Many deontologists are pluralists who believe that the value of the consequences of our actions is often one important consideration when deciding what to do, and this seems to be the most attractive type of deontological theory. As Rawls (1971, p. 30) put it, "All ethical doctrines worth our attention take consequences into account in judging rightness. One which did not would simply be irrational, crazy."

## 1.2 Why the question is important

My main question in this dissertation is whether the rough future-shaping argument works. This is an important question because if it does work, it has significant practical implications. It's also important, and theoretically interesting, because properly addressing it illuminates a number of fundamental theoretical issues. Finally, it's important because despite the overwhelming importance of its subject matter—after all it addresses the real, even if remote, risk of the total destruction of human civilization—it has, strangely, been largely neglected by moral philosophers. Moreover, my question has not been given a satisfactory answer, and it has not given the careful, lengthy, systematic analysis which a question of this importance cries out for. My aim in this thesis is to rectify that lack.

### 1.2.1 Why the question is practically significant

People currently evaluate good accomplished mostly by looking at proximate benefits. But if the rough future-shaping argument works, what matters most is how our actions affect the far future. It is unclear how closely these two standards overlap, but they may significantly diverge in some cases. If they greatly diverge, it may mean that this misunderstanding could cause us to do much less good than we otherwise would. We may be able to get a handle on how much the goals diverge if we investigate various empirical questions, and this is very important if my arguments are correct.

To elaborate on this idea, it helps to distinguish between very *broad* ways of trying to shape the far future—such as trying to improve education for talented youth—and very *targeted* ways of shaping the far future—such as trying to prevent an asteroid from hitting the Earth. And, of course, there is a spectrum between them, where options like, "Teach our children to be good stewards of the future," are somewhere in between. The general distinction is that broad approaches focus on unforeseeable benefits from ripple effects, whereas targeted approaches aim for more specific effects on the far future, or aim at a relatively narrow class of positive ripple effects. If we accept the rough future-shaping argument and we continue to evaluate good accomplished using ordinary standards, we are essentially betting that ripple effects from everyday good works have positive consequences for the far future that exceed the benefits of our best targeted attempts. Once one accepts the rough future-shaping argument, it becomes important to consider exactly where on the broad/targeted spectrum the best opportunities for shaping the far future lie. The more targeted the best opportunities are, the more practically significant my arguments become.

Currently no one knows how extensively and/or positively ripple effects of ordinary actions affect the far future. And consequently, no one knows exactly how broad or narrow the best ways of shaping the far future are. And therefore, no one knows whether ordinary standards for evaluating outcomes are majorly mistaken. This seems like a major gap in knowledge which, as far as I know, is not being seriously and carefully investigated, and the gap becomes clear if my arguments are accepted.

If we believe that the best ways of shaping the far future are relatively broad, it becomes important to find out what kinds of causes have the most significant ripple effects of the right kind. It may be important to identify causes that have a high ratio of far future benefits to proximate benefits, since these causes are likely to be much more important than they are by ordinary standards. In some cases, we know that our actions have relatively limited ripple effects on the far future, and this can lead to surprising moral conclusions. For instance, I believe that preventing animal suffering in factory farms has very little ripple effects on the far future. By ordinary standards though—at least

by ordinary non-speciesist standards—reducing suffering in factory farms would be extremely important. But because I accept the future-shaping argument and I believe reducing suffering on factory farms does very little to shape the far future, I now consider reducing suffering on factory farms to be much less important than I did previously. Or, more precisely, I now consider reducing suffering on factory farms to be much less important *in comparison with other things* than I did previously. To take another example, saving lives in poor countries may have significantly smaller ripple effects than saving and improving lives in rich countries. Why? Richer countries have substantially more innovation, and their workers are much more economically productive. By ordinary standards—at least by ordinary enlightened humanitarian standards—saving and improving lives in rich countries is about equally as important as saving and improving lives in poor countries, provided lives are improved by roughly comparable amounts. But it now seems more plausible to me that saving a life in a rich country is substantially more important than saving a life in a poor country, other things being equal.

If we believe that the best ways of shaping the far future are likely to be relatively targeted, it becomes important to ask various other questions whose answers are currently unknown. These questions include:

1. What are the most significant existential risks (including significant global threats which may indirectly lead to existential catastrophes)? Are we prepared for them? Can we be more prepared for them at bearable costs?

2. If we can be more prepared for existential risks, what are the most effective targeted approaches for being more prepared?

3. Apart from existential risks, are there other significant potential trajectory changes which are suited to a targeted approach? What are the most promising ones?

4. How are proximate benefits, speeding up development, existential risk, and other trajectory changes related to each other?

5. Is the expected value of the future negative? Some serious people—including Parfit (2011, Volume 2, chapter 36), Williams (2006), and Schopenhauer (1942)—have wondered whether all of the suffering and injustice in the world outweigh all of the good that we've had. I tend to think that our history has been worth it, that human well-being has increased for centuries, and that the expected value of the future is positive. But this is an open question, and stronger arguments pointing in either direction would be welcome.

In this dissertation, I don't address these empirical questions at any length. Though many of them would be very challenging to speculate about, they would be very valuable to answer if my conclusions are correct and they have received relatively little systematic attention.

### 1.2.2 Why the question is theoretically interesting

Assessing the rough future-shaping argument is theoretically interesting because it requires grappling with some basic issues in population ethics and decision theory, and there are some interesting connections between population ethics and decision theory which are illuminated by thinking about these arguments.

It is fairly obvious why the future-shaping argument has to engage with population ethics: the argument says that it would be very, very good if there were many future generations, and whether that is true depends on questions of population ethics. On a straightforward total utilitarian view, provided that future generations would have good lives, it would obviously be very good if there were many future generations. As we'll see later, there is a more general class of views which have similar implications. But for now, it suffices to observe that there are two major ways we could change the total utilitarian view which would substantially affect the value of ensuring the existence of many future generations. Total utilitarians determine the value of an alternative by adding up the total well-being of all the people that would ever exist if that alternative were chosen. The first kind of revision involves tinkering with the *form* of the aggregation procedure (e.g. changing from summing over people to taking averages, giving additional people diminishing marginal value, or doing something much more complicated). The second kind of revision involves tinkering with the *scope* of the aggregation procedure (changing from aggregating over all people that would ever exist if the alternative were chosen to aggregating over some subset of those people). The idea behind restricting the scope of the aggregation is to allow us to ignore (or discount) any supposed non-instrumental reasons to create "extra" people. Person-Affecting Views are the most famous and common examples of the second type of view.

According to standard Person-Affecting Views, there can be no opportunity cost in failing to create a person because if that person is never created there is never a person who could have benefited from being created. On this view, the fact that a person's life would go well if he lived could not imply that creating this person would be in some way good. Furthermore, the only reason it would be important to positively shape the far future is that it would affect the interests of people alive today. If this view is correct, then the rough future-shaping argument straightforwardly fails,

so defenders of this argument must address Person-Affecting Views.

If we hold that it is good for there to be future generations and that each additional generation is equally important (other things being equal), then it is very plausible that, as the rough future-shaping argument assumes, it would be very good if there were many future generations. But if we hold instead that once there have been "enough" good lives, additional good lives count for less, then some ways of shaping the far future, such as reducing existential risk, may be much less important. If we want to thoroughly address the rough future-shaping argument, we have to deal with issues like this as well.

On the decision theory side, there is substantial theoretical interest in untangling a number of messy issues associated with thinking about murky, low-probability, high-stakes risks. One reason for this is that it is very hard to create positive trajectory changes or reduce existential risk, and thinking about the value of pursuing these goals requires us to consider some murky, low probability scenarios. If we are willing to use speculative expected value calculations to reach the conclusion that shaping the far future is overwhelmingly important, why don't we instead decide to spend our resources on even more extreme long shots, such as aiming for infinitely good outcomes? Thinking about this issue raises classical theoretical issues from decision theory—such as what to say about Pascal's Wager and gambles with infinite expected value, as well as gambles with enormously high finite expected value—but it raises them in a more realistic context.

Furthermore, these issues in decision theory are, to some extent, entwined with issues in population ethics. Many views in population ethics, including the views which could most straightforwardly be used to support the rough future-shaping argument, implicitly claim that there is no upper limit to how good outcomes could be. For example, total utilitarianism implies that there is no upper limit to how good outcomes can be. It implies this because for any state of the world, adding an additional person with a given level of happiness makes the outcome better by a fixed amount, and if we just add sufficiently many such people, the value of the outcome can exceed any given amount. Similar arguments show that many other theories in population ethics—including prioritarian and critical-level forms of utilitarianism,[3] as well as pluralistic theories which put no upper limit on the potential importance of utilitarian considerations—have similar implications. Other views in population ethics imply that there is an upper limit to how good outcomes can be. For example, some average utilitarians may believe that there is an upper limit to how well any individual's life can go

[3]Here by *critical-level utilitarianism* I mean the type of theory developed by Broome (2004) and Blackorby et al. (2005), and not the sort of two-level utilitarianism advocated by Hare and others. Critical-level utilitarianism is a theory of population ethics according to which there is some positive threshold $c$ such that the value of a world equals the sum of (the first person's well-being minus $c$, the second person's well-being minus $c$,..., the last person's well-being minus $c$).

for him. Given that assumption, there is an upper limit to the average level of well-being. On some Person-Affecting Views, there is an upper limit to how good outcomes can be. Why? Some views in this category may hold that there will only be finitely many people whose interests matter for the goodness of outcomes, and there may be an upper limit to how well off each of these people could be, which would imply that there would be an upper limit to how good outcomes could be.

Later in this dissertation, I argue that if there is no upper limit to how good outcomes could be and some outcomes could be infinitely good, then infinite considerations essentially dominate decision-making under uncertainty, at least under some standard assumptions from decision theory. For example, I argue that if we take some theory like total utilitarianism, which implies that there is no upper limit to how good an outcome could be, and that theory says that some outcomes are infinitely good, such as an outcome with infinitely many happy people or an outcome where one person is happy forever, then according to those theories, the best decision under uncertainty is almost entirely determined by considerations like

- "What would maximize the probability of having infinitely many happy people?"
- "What would create a 'larger' expected infinite amount of happiness?"
- "What would minimize the probability of having infinitely many people that are unhappy?"
- etc.

rather than ordinary considerations like

- "Which of these global health programs would save the most lives?"
- "Which of these policies would maximize GDP?"
- "Which of these lesson plans would maximize student achievement?"
- etc.

As with minimizing existential risk or shaping the far future more generally, the best approaches to pursuing these improbable goods may be very broad. Therefore, the practical implications in terms of *what* it would be best to do may not differ greatly from common sense. However, what we believe about *why* it would be best to pursue one strategy rather than another may change dramatically. It would be strange if the ultimate determinant of the value of some low-income housing program was basically a function of the probability that it eventually produced some infinitely valuable outcome.

What this shows is that there is a deep connection between population ethics and decision theory. Many theories of population ethics I'm aware of have some implausible and seemingly unintended implications about how it would be best to make decisions under uncertainty. Because there is this connection between our two subjects, and both issues naturally arise when dealing with the rough future-shaping argument, analyzing the rough future-shaping argument can help illuminate population ethics and decision theory.

### 1.2.3 Some reasons the question has not been given a satisfactory answer

It is non-obvious whether the rough future-shaping argument works, and it matters a great deal whether it works. Have we adequately addressed the question? I believe not. The main reason I believe this is because I searched the literature in philosophy and economics to see if the question had been given a satisfactory answer and found limited discussion of the issues, and nothing approaching consensus.

John Broome, who has written extensively on both population ethics and the ethics of climate change, has made some comments which suggest that he has a similar view:

> The Intergovernmental Panel on Climate Change reports several studies of how global temperatures will increase in the long run if atmospheric greenhouse gases reach the warming equivalent of about 550 parts per million of carbon dioxide (a level expected within a few decades). Most of the studies estimate the probability is 5 percent or more that the increase will be above eight degrees Celsius (14.4 degrees Fahrenheit). The disruption caused by such temperatures would pose some risk—no one can say how much—of a devastating collapse of the human population, perhaps even to extinction. Any such event would be so bad that even multiplied by its small chance of occurrence, its badness could dominate all calculations of the harm that climate change will cause. Working out how bad such an event would be is an urgent but very difficult ethical problem. (Broome, 2008, p. 72)

> ... we cannot just assume the small chance of catastrophe is the most important thing about climate change. It may or may not be. To know which, we must work out just how bad the catastrophe would be. ... working this out will be difficult. Extreme climate change will certainly cause a collapse of the human population. It may cause the extinction of humanity. Naively, we think of these as terrible disasters. Perhaps they are. But to know whether they are terrible, and if they are, how terrible, we must investigate

> how good or bad it is for a person to exist. This is a difficult task for moral philosophy. (Broome, 2010, p. 116)

Broome's comments suggest that he agrees that this problem is important and unanswered. This is both a good sign and a bad sign for someone hoping to gain insight by thinking hard about the argument. It's a good sign because if we make significant progress, then we'll have some original insights. It's a bad sign because it may be that people have steered clear of analyzing the argument for good reasons. So it helps to consider why the argument hasn't been fully addressed.

Since we don't know how good the rough future-shaping argument is and it potentially matters a lot how good it is, it would seem that best reason not to analyze the argument would be that it would be too hard to make progress on analyzing the argument. There are a lot of good reasons to think it would be hard to make progress analyzing this argument. First, evaluating the argument requires making judgments about the distant future. We have to say something about how likely various civilization-destroying catastrophes would be, how much we could reduce the risk of these catastrophes, and how long humanity could be expected to survive if we avoid these catastrophes. We can't hope for much accuracy in any of these estimates. Second, evaluating the argument requires resolving challenging issues in population ethics and decision theory. We know from the work of Parfit (1984), Temkin (2012), Broome (2004), and Arrhenius (2013) that population ethics is riddled with paradoxes. And we also know from the work of Hajek (2003) and Bostrom (2009) that developing a version of decision theory which handles murky, low-probability, high-stakes gambles is no picnic either. Third, to come to reasonable judgments about the value of shaping the far future, we'd not only have to come up with good enough views about the far future, population ethics, and decision theory, but also manage to compare shaping the far future with conventional ways of doing good.

On the other hand, there are reasons that academics would steer clear of analyzing this type of argument that have little to do with the potential value of making progress on the argument. One of the most significant reasons is that the topic of shaping the far future doesn't neatly fit into any particular academic field, and good questions which don't neatly fit into any particular academic field can get neglected for reasons unrelated to the importance of the questions. Assessing the value of reducing existential risk involves speculating about how likely humans are to be wiped out by different types of catastrophes (such as nuclear wars, asteroids, catastrophic climate change, etc.), speculating about how long humans could potentially survive in the distant future, and doing population ethics and decision theory to say how good it would be to try to prevent these disasters.

Philosophers may enjoy some amount of speculation about more conceptual and theoretical issues in science (such as the nature of perception or the nature of economic explanations), but this kind of speculation isn't standard fare for philosophers. And most of the people who have some expertise in these specific domains would be unlikely to engage with the basic issues of moral philosophy. Together, these factors leave us with a good question that isn't getting fully addressed by any group of academics.

In summary, this question does not seem to be adequately addressed, and that may partially be because the question is very hard to answer, but it may partially be because the question straddles multiple fields in a way that may lead to neglect of the question for reasons unrelated to the importance of the question.

## 1.3 My approach to answering the question

My bet is that whether the rough future-shaping argument works is largely a philosophical question, and that we can make significant progress on it by doing a minimal amount of empirical background setting, finding the critical normative assumptions upon which the argument turns, and then assessing the plausibility of those assumptions. Crucially, we do not need to "solve" population ethics in order to address the question. Instead, I argue, it suffices to defend some plausible, more minimal assumptions. This is the approach I take in chapter 3.

### 1.3.1 What are the critical philosophical assumptions behind the rough future-shaping argument?

In trying to estimate the value of the future, I focus on devising some approximation techniques that should work in cases that we could potentially encounter, rather than trying to "solve" some part of population ethics in a precise way that works in all possible cases. The basic idea is to divide the history of the world into periods of time of some large duration (such as 100, 1000, or 10,000 years) and then say how the value of the whole (all of history) is determined by the value of the parts (what happens in different periods of history). When I talk about "periods," I mean these chunks of time. I make very minimal assumptions about how the value of a single period is determined, and try not to worry about whether the my approximation technique is perfectly accurate or fails in edge cases.

Obviously, you might get different answers depending on how you carve up the world into periods of time, and there is no "privileged" way of carving up history into periods. This does not concern

me because I'm working out an approximation technique, rather than trying to provide a "final theory" of population ethics. Using an approximation technique requires some judgment and often requires making some arbitrary decisions, so it is not problematic if there is some arbitrariness in how the technique is applied in particular cases. The approximation could be improved in various ways, such as by taking an average over different ways of doing the approximation, but again this is not very important for the arguments I make.

In chapter 3, I explain this approach in greater detail. The major normative assumptions behind the approach are:

> *Additionality*: If "standard good things" happen during a period of history—there are people, the people have good lives, society is organized appropriately, etc.—that makes that period go better than a period where nothing of value happens
>
> *Period Independence*: By and large, how well history goes as a whole is a function of how well things go during each period of history; when things go better during a period, that makes the history as a whole go better; when things go worse during a period, that makes history as a whole go worse; and the extent to which it makes history as a whole go better or worse is independent of what happens in other such periods.[4]
>
> *Temporal Impartiality*: The value of a particular period is independent of when it occurs.
>
> *Risk Neutrality*: The value of an uncertain prospect equals its expected value.

If these assumptions are true, I argue, we can quantify the additional value of additional periods of history by looking at how well additional periods go and adding that to a running total of how well everything has gone so far. That way, if we have a very long future that is about equally good in each of its periods, the value of that future will be roughly proportional to how long it lasts, and we can say this without making major assumptions about what makes a period of history go well. Then, if we can increase the probability that this happens, that would have a lot of expected good, which means it would be very important to do that.

Chapter 3 does some work to illustrate the plausibility of these assumptions, and also to respond to some preliminary objections. Later chapters discuss some of the more problematic assumptions, and important challenges to them. For example, later chapters discuss the plausibility of appealing

[4] I do not mean *causally* independent, I mean *metaphysically* independent. If something happens in a person's life at some point, there are two ways in which how much that makes his life go better may depend on what happens at other times in his life. It may just cause different events to happen later, which makes his life go better in a straightforward, causal way. Or, it may not have important causal impacts, but it might make his life fit into a more meaningful story and thereby make it better—that's a sort of metaphysical dependence. Period Independence is denying that there is that second sort of metaphysical dependence for the world in general across long periods of time.

to a Person-Affecting View to resist Additionality, or claiming that additional lives have diminishing marginal value as a way of resisting Period Independence.

### 1.3.2 My methodology for assessing the plausibility of these assumptions

Chapter 2 is about methodology for normative and applied ethics. Producing a chapter on this topic was called for because I am considering arguments for a very contrarian view, and there's a burden on me to say something about how people could possibly be as wrong as the rough future-shaping argument says they are. Second, I have some non-standard views about how seriously we should take our pre-theoretic moral judgments, this shows up to some extent in my arguments, and this approach needs to be defended.

In short, I argue that our moral judgments are less reliable than many would hope, and this has specific implications for methodology in normative ethics. Three sources of evidence indicate that our intuitive ethical judgments are less reliable than we might have hoped: a historical record of accepting morally absurd social practices; a scientific record showing that our intuitive judgments are systematically governed by a host of heuristics, biases, and irrelevant factors; and a philosophical record showing deep, probably unresolvable, inconsistencies in common moral convictions. I argue that this has the following implications for moral theorizing: we should trust intuitions less; we should be especially suspicious of intuitive judgments that fit a bias pattern, even when we are intuitively confident that these judgments are not a simple product of the bias; we should be especially suspicious of intuitions that are part of inconsistent sets of deeply held convictions; and we should evaluate views holistically, thinking of entire classes of judgments that they get right or wrong in broad contexts, rather than dismissing positions on the basis of a small number of intuitive counterexamples. In addition, I argue that many of the specific biases that I discuss would lead us to predict that people would, in general, undervalue most of the available ways of shaping the far future, including speeding up development, existential risk reduction, and creating other positive trajectory changes.

### 1.3.3 What are the major conclusions about population ethics? What kind of arguments are offered in favor of these conclusions?

In chapters 4 and 5, I argue against two of the strongest objections to the rough future-shaping argument. These objections come from population ethics, and are based on Person-Affecting Views and views according to which additional lives have diminishing marginal value.

In chapter 4, I argue against appealing to Person-Affecting Views as a response to rough future-

shaping argument for the following reasons:

1. The main theoretical (contrasted with intuitive) arguments for this position are unconvincing.

2. Person-Affecting Views have many counterintuitive implications about particular cases. The worst of these implications involve the permissibility of having children and the desirability of ending civilization.

3. Some of the important theoretical motivations for adopting a Person-Affecting View can be captured without appealing to a Person-Affecting View.

4. The weaker, more plausible versions of the Person-Affecting View do not undermine the rough future-shaping argument.

In chapter 5, I consider the objection that additional periods of history have diminishing marginal value, so that after we have had "enough" good periods of history, it matters less whether there are additional good periods of history in the future. To get a grip on the objection, imagine that humans survive the next 1000 years, and their lives go well. How good would it be if they survived for another thousand years, with the same or higher quality of life? What if they survived another thousand years beyond that? Consider three kinds of answer:

1. The Period Independence answer: It would be equally as important in each such case.

2. The Capped Model answer: After a while, it gets less and less important. Moreover, there is an upper limit to how much value you can get in this way.

3. The Diminishing Value answer: After a while, it gets less and less important. However, there is no upper limit to how much value you can get in this way.

I consider the costs and benefits of each type of answer, and argue that Period Independence is the most plausible. I argue that Period Independence does a better job of capturing the following intuitions: the intuition that it is always bad to miss a good future that we could have had, the intuition that one does not need to know how long we've flourished in the past to say how good it would be to have additional good periods of history, and the intuition that it would not be bad to give humanity an additional period of flourishing when the potential downsides of doing this are small. I also argue that even if, intuitively, creating additional people at a time when many people exist is not very valuable, there is no good argument from this intuition to the conclusion that creating additional people has little value at times when no people would otherwise exist.

### 1.3.4 Long shots, upper limits to the value of outcomes, and infinities: a theoretical challenge for my arguments

In chapters 6 and 7, I discuss a major theoretical challenge to my formalization of the rough future-shaping argument. I call it a "theoretical challenge" rather than just a "challenge" because my intuition is that if my formalization would have worked but for this difficulty, it should be possible to create another formalization that would work if this difficulty were dealt with. However, that is not obvious, and it is not clear what the best way to formalize the rough future-shaping argument is once the theoretical challenge has been appreciated.

The entry point for the discussion is to observe that if there are opportunities for individuals to creating positive trajectory changes, they probably only work with small probability. This is especially true in the case of existential risk reduction. Thus, the rough future-shaping argument asks us to be happy with having a very small probability of averting an existential catastrophe, on the grounds that the expected value of doing so is extremely enormous, even though there are more conventional ways of doing good which have a high probability of producing very good, but much less impressive, outcomes. Essentially, we're asked to choose a long shot over a high probability of something very good. In extreme cases, this can seem irrational on the grounds that it's in the same ballpark as accepting a version of Pascal's Wager.

In chapter 6, I make this worry more precise and consider the costs and benefits of trying to avoid the problem. When making decisions under risk, we make trade-offs between how good outcomes might be and how likely it is that we get good outcomes. There are three general kinds of ways to make these tradeoffs. On two of these approaches, we try to maximize expected value. On one of the two approaches, we hold that there are limits to how good (or bad) outcomes can be. On this view, no matter how bad an outcome is, it could always get substantially worse, and no matter how good an outcome is, it could always get substantially better. On the other approach, there are no such limits, at least in one of these directions. Either outcomes could get arbitrarily good, or they could get arbitrarily bad. On the third approach, we give up on ranking outcomes in terms of their expected value.

The main conclusion of chapter 6 is that all of these approaches have extremely unpalatable implications. On the approach where there are upper and lower limits, we have to be *timid*—unwilling to accept extremely small risks in order to enormously increase potential positive payoffs. Implausibly, this requires extreme risk aversion when certain extremely good outcomes are possible and extreme risk seeking when certain extremely bad outcomes are possible, and it requires making one's

ranking of prospects dependent on how well things go in remote regions of space and time.

In the second case, we have to be *reckless*—preferring very low probabilities of extremely good outcomes to very high probabilities of less good, but still excellent, outcomes—or rank prospects non-transitively. I then show that, if a theory is reckless, what it would be best to do, according to that theory, depends almost entirely upon what would be best in terms of considerations involving infinite value, no matter how implausible it is that we can bring about any infinitely good or bad outcomes, provided it is not *certain*. In this sense, there really is something deeply Pascalian about the reckless approach.

Some might view this as a reductio of expected utility theory. However, I show that the only way to avoid being both reckless and timid is to rank outcomes in a circle, claiming that $A$ is better than $B$, which is better than $C$,..., which is better than $Z$, which is better than $A$. Thus, if we want to avoid these two other problems, we have to give up not only on expected utility theory, but we also have to give up on some very basic assumptions about how we should rank alternatives. This makes it much less clear that we can simply treat these problems as a failure of expected utility theory.

What does that have to do with the rough future-shaping argument? The problem is that my formalization of the rough future-shaping argument commits us to being reckless. Why? By Period Independence, additional good periods of history are always good, how good it is to have additional periods does not depend on how many you've already had, and there is no upper limit (in principle) to how many good periods of history there could be. Therefore, there is no upper limit to how good outcomes can be. And that leaves us with recklessness, and all the attendant theoretical difficulties.

At this point, we are left with a challenging situation. On one hand, my formalization of the rough future-shaping argument seemed plausible. However, we have an argument that if its assumptions are true, then what it is best to do depends almost entirely on infinite considerations. That's a very implausible conclusion. At the same time, the conclusion does not appear to be easy to avoid, since the alternatives are the so-called "timid" approach and ranking alternatives non-transitively.

In chapter 7, I discuss how important it would be to shape the far future given these three different possibilities (recklessness, timidity, and non-transitive rankings of alternatives). As we have already said, in the case of recklessness, the best decision will be the decision that is best in terms of infinite considerations. In the first part of the chapter, I highlight some difficulties for saying what would be best with respect to infinite considerations, and explain how what is best with respect to infinite considerations may depend on whether our universe is infinitely large, and whether it makes sense to say that one of two infinitely good outcomes is better than the other.

In the second part of the chapter, I examine how a timid approach to assessing the value of

prospects bears on the value of shaping the far future. The answer to this question depends on many complicated issues, such as whether we want to accept something similar to Period Independence in general even if Period Independence must fail in extreme cases, whether the universe is infinitely large, whether we should include events far outside of our causal control when aggregating value across space and time, and what the upper limit for the value of outcomes is.

In the third part of the chapter, I consider the possibility of using the reckless approach in contexts where it seems plausible and using the timid approach in the contexts where it seems plausible. This approach, I argue, is more plausible in practice than the alternatives. I do not argue that this mixed strategy is *ultimately correct*, but instead argue that it is the best available option in light of our cognitive limitations in effectively formalizing and improving our processes for thinking about infinite ethics and long shots.

## 1.4 What are the major conclusions of this dissertation? What are the remaining questions?

In summary, the major conclusions of this dissertation are:

1. There is a very plausible argument that shaping the far future is overwhelmingly important.
2. Person-Affecting Views face extreme difficulties and cannot plausibly be used to rebut this argument.
3. Views according to which additional future people have diminishing marginal value have some very significant problems and cannot be used to rebut this argument.
4. What appears to be the best argument for the overwhelming importance of shaping the far future is caught up in a challenging paradox about how to make decisions when there are very small probabilities of enormously good outcomes and infinitely good outcomes.
5. Until we have an adequate resolution to this paradox, if there is one, in practice, we should continue to rely on the rough future-shaping argument and generally ignore extremely far-fetched enormously good outcomes and infinitely good outcomes.

Some of the remaining questions are primarily questions for normative ethics. They include:

1. Are there upper and lower limits to how valuable outcomes could be?

2. If there are upper and lower limits to how valuable outcomes could be, is Period Independence still approximately correct in situations we could foreseeably face, or do we have to rethink the value of the future in a much more significant way?

3. If there are not upper and lower limits to how valuable outcomes could be, how should we compare prospects involving infinitely valuable outcomes?

4. How should we compare infinitely valuable outcomes with other infinitely valuable outcomes?

5. Is there any clear improvement on using the reckless approach in the contexts where it delivers plausible conclusions and using the timid approach in the contexts where it delivers plausible results?

I begin to address these questions in chapters 6 and 7, but they are very challenging, and there is much more work to be done.

Other significant remaining questions are largely empirical. They include all the questions I identified in Section 1.2.1.Very little has been done to answer these empirical questions. If my conclusions are even roughly correct, then it would be very valuable to make progress on these questions. In my view, it is likely that we could make significant progress on at least some of these questions. Hopefully some people will try.

# Chapter 2

# How Could We Be So Wrong?

## Introduction

In this chapter I argue that our moral judgments are subject to systematic errors that are hard to detect and hard to correct, and I draw some methodological conclusions in light of this.

In the first section, I describe a Bayesian framework for moral philosophy, using curve fitting in philosophy of science as my model for reasonable epistemic inquiry. In the second section, I use this framework to explain how information about our systematic errors should affect our methodology in normative ethics. I argue that:

1. We should give relatively less weight to fit with intuition as a criterion of theory choice in moral philosophy.

2. We should evaluate views holistically, thinking of entire classes of judgments that they get right or wrong in broad contexts, rather than dismissing positions on the basis of a small number of intuitive counterexamples

3. We should give relatively more weight to background meta-ethical theories, our basic hunches about which moral theories are likely to be true, and our basic epistemic standards, perhaps including simplicity.

The next three sections point to the evidence of moral error.

I point to three sources of evidence for the existence of widespread error that is hard to detect and hard to correct: (i) a historical record of accepting morally absurd social practices; (ii) a scientific record showing that our intuitive judgments are systematically governed by a host of heuristics,

biases, and irrelevant factors; and (iii) a philosophical record showing deep, probably unresolvable, inconsistencies in common moral convictions. (In making this claim about moral error, I do not distinguish between judgments about cases and judgments about general principles. Both are less reliable than most believe.) A special conclusion from examining the heuristics and biases literature is that we should be especially suspicious of moral judgments which could be explained by biased heuristics, even when we believe that the judgments were not the result of such heuristics. A special conclusion from looking at the philosophical record is that developing a satisfactory theory of population ethics should mostly be an exercise in damage control—we should try to find theories that capture our values where we can, but we should expect major revisions in the end—rather than an attempt to find a theory that accords with all of our deepest convictions.

In the sixth section, I argue that systematic error also makes it less surprising that few people have accepted the conclusion of the rough future-shaping argument: that trying to shape the far future is overwhelmingly important—more important than almost anything which would not have a significant effect on the far future. Specific biases that I have discussed, together with a simple understanding of how heuristic-based reasoning works, strongly suggest that intuition would dramatically underestimate the value of shaping the far future.

I owe some of they key insights in this chapter to Robin Hanson, who first alerted to the similarities between the curve-fitting problem in philosophy of science and methodology in moral philosophy.[1]

## 2.1 Curve fitting and the significance of widespread correlated error

### 2.1.1 Introduction to curve fitting

Curve fitting is a problem frequently discussed in the philosophy of science. In the standard presentation, a scientist is given some data points, usually with an independent variable and a dependent variable, and is asked to predict the values of the dependent variable given other values of the independent variable. Typically, the data points are *observations*, such as "measured height" on a scale or "reported income" on a survey, rather than true values, such as height or income. Thus, in making predictions about additional data points, the scientist has to account for the possibility of error in the observations. By an *error process* I mean anything that makes the observed values of

[1] See the methodological points discussed in Hanson (2002, pp. 154-157).

the data points differ from their true values. Error processes could arise from a faulty scale, failures of memory on the part of survey participants, bias on the part of the experimenter, or any number of other sources. While some treatments of this problem focus on predicting observations (such as measured height), I'm going to focus on predicting the true values (such as true height).

Famously, there are trade-offs between simplicity and fit with data in the curve-fitting problem. In short, you can always maximize fit or maximize simplicity, but typically neither yields acceptable results, so one has to make trade-offs between the two. In the graph below, we have data points (the diamonds) and two curves which could be used to predict the values of $Y$ given additional observations of $X$.

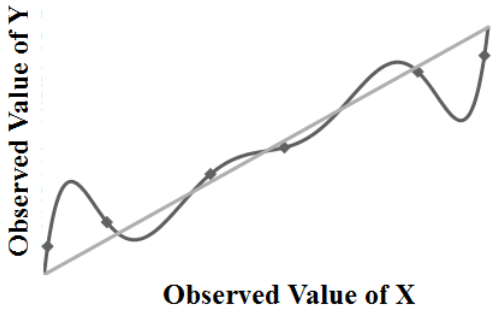


Figure 2.1: Curve-fitting Example

For any consistent data set, it is possible to construct a curve that fits the data exactly, as I've done with the silly 6th order polynomial. In fact, there are infinitely many funny polynomial curves that could fit all of the observations exactly. If the scientist chooses one of these polynomial curves for predictive purposes, the result will usually be *overfitting*, and the scientist will make worse predictions than he would have if he had chosen a curve that did not fit the data as well, but had other virtues, such as a straight line. On the other hand, always going with the simplest curve and giving no weight to the data leads to *underfitting*. Identifying the best way to do this, and saying why it's the best way to do this, is a familiar problem in philosophy of science (Baker, 2011).

I intend to carry over our thinking about curve fitting in science to reflective equilibrium in moral philosophy, so I should note immediately that curve fitting is not limited to the case of two variables. When we must understand relationships between multiple variables, we can turn to multiple-dimensional spaces and fit planes (or hyperplanes) to our data points. Different axes might correspond to different considerations which seem relevant (such as total well-being, equality,

number of people, fairness, etc.), and another axis could correspond to the value of the alternative, which we can assume is a function of the relevant considerations. Direct Bayesian updating on such data points would be impractical, but the philosophical issues will not be affected by these difficulties.

I should also note that when I carry over conclusions, my argument won't rest on an *analogy*. I'll be saying that the analysis carries over literally.

### 2.1.2 The Bayesian approach to curve fitting

On a Bayesian approach to this problem, the scientist would consider a number of different hypotheses about the relationship between the two variables, including both hypotheses about the phenomena (the relationship between $X$ and $Y$) and hypotheses about the error process (the relationship between observed values of $Y$ and true values of $Y$) that produces the observations. The Bayesian approach can be instructively outlined as in the diagram below:

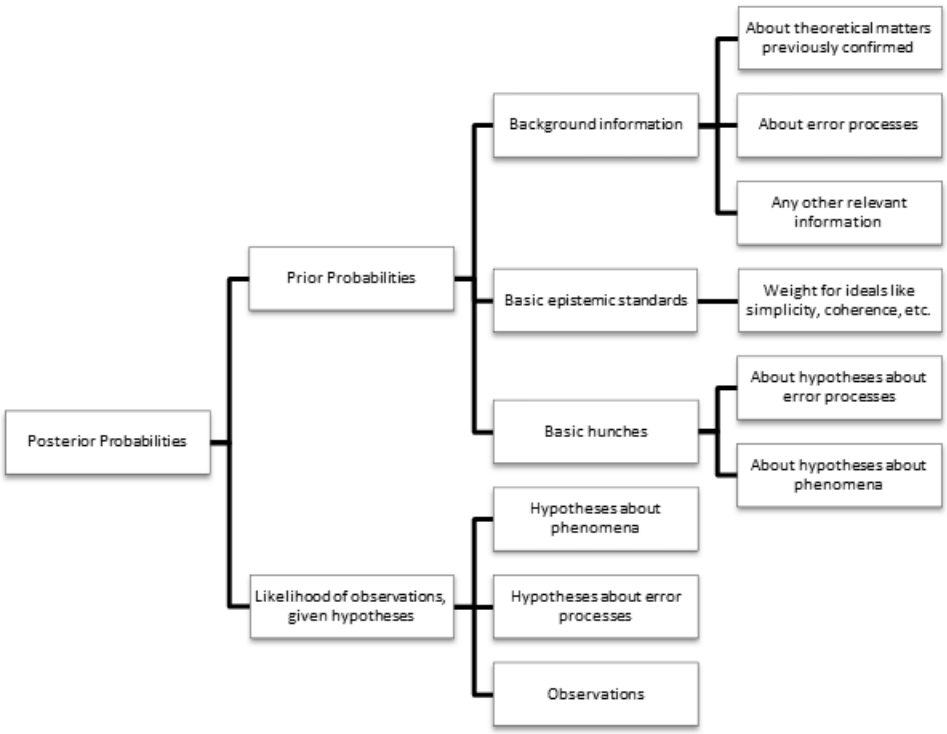


Figure 2.2: The Bayesian approach to curve-fitting

In this diagram, each parent box is determined by the boxes that are its children. So posterior probabilities are determined by priors and the likelihoods of the observations given different hypotheses;

likelihoods are determined by observations, hypotheses about error processes, and hypotheses about the phenomena; priors are determined by background information, basic epistemic standards, and basic hunches; and so forth. I'll explain more about what these technical terms mean briefly.

Starting from the left, our *posterior probabilities* are our judgments at the end of the investigation about how likely the different hypotheses we're considering are. The posterior probabilities are determined by two scores that each hypothesis receives, both of which are probabilities. The first is *likelihood*, and it depends on what observations we had. The likelihood of some observations, relative to a hypothesis, is just the probability of observing exactly those observations if the hypothesis in question were true. The second score is *prior probability*, which is a measure of how plausible the hypothesis was in advance of getting the new data points. The prior probability includes all of the available information, apart from the new data. These scores are multiplied together and then renormalized to determine posterior probabilities. The likelihoods may be challenging to construct in practice but they do not raise philosophically vexing issues—they are determined uniquely by the hypotheses (which model both the phenomena and the error process) and the observations.

The prior probabilities, on the other hand, cannot be mechanically produced, and call for judgment. To some extent, the prior probabilities will depend on *background information* about the system in question. This can include theoretical information about matters previously confirmed, information about the error processes, or any other relevant information. For example, if the scientist is making predictions about the relationship between time to finish a triathlon and time to finish a marathon, the scientist will know that the best prediction curve will always slope upwards (given the background knowledge that people who are good at running tend to be better at triathlons). The scientist might have information about an error process in his measurement; perhaps the timer used was discovered to have some significant random error.

But the priors will also depend on the scientist's *basic epistemic standards*, such as how much weight the scientist puts on theoretical virtues like simplicity, elegance, explanatory power, limited ontology, etc. Many philosophers of science, and many scientists, would recommend using priors that favor simpler hypotheses.[2] And, in practice, it seems that most cases of overfitting are caused by giving too little weight to simplicity. Here, complexity is usually thought of as choosing too many parameters for one's curve. On this way of understanding simplicity, a 3rd degree polynomial is more complex than a 2nd degree polynomial, which is more complex than a 1st degree polynomial.

Finally, the priors will depend (in a significant way) on *basic hunches*, pure guesses about which hypotheses about the phenomena and the error processes are more likely than which other hypothe-

[2]Baker (2011), Carnap (1950), and Swinburne and Ebrary (1997) are examples that readily come to mind.

ses. An example of a basic hunch might be the assumption that things that taste really, really bad are not good for your health.[3]

There are deep questions about prior probabilities that I am not going to address. For example, I'm not going to say anything about what restrictions there might be on which basic hunches and/or basic epistemic standards are rationally acceptable. And I'm not going to try to recommend a specific set of basic epistemic standards/basic hunches. Suffice it to say that appealing to fit with data alone is not feasible, and some theoretical factors and/or basic hunches, have to get used in order to make predictions on the basis of data. Others have argued for the plausibility of the Bayesian approach to the curve fitting problem, and now is not the time to recount those arguments. Instead, I will simply assume that this is the correct way to think about the problem of curve fitting, and draw out the implications.

### 2.1.3 When you expect more error, rely on priors more

Since the prior probabilities range over all hypotheses about the situation (including both the model of the phenomenon itself and the error process), the prior probabilities include the scientist's expectations about how reliable his measurements are. Here is an important conclusion: the more error the scientist expects, the more weight he should place on his priors, relative to fit with the data. In these cases, the scientist should be more willing to chalk up observations to error, and should rely more on background information about the hypotheses about the phenomena, his basic hunches, and his basic epistemic standards.

This claim can get some intuitive plausibility by testing the limiting cases. If the scientist thinks that his observation process has *no error at all*, (e.g., God is telling him the answers) then he can automatically reject any hypothesis that does not exactly fit his observations. If the scientist thinks that his observation system is totally error prone (e.g., his measuring device outputs "6" regardless of the state of the world), then he depends entirely on his prior beliefs. In intermediate cases, he will place intermediate weight on fit and prior beliefs.

[3]Some might question the distinction between basic epistemic standards and basic hunches, contending that all basic epistemic standards are merely generalized basic hunches. Nothing I'm going to say will turn on whether these things belong in separate categories.

### 2.1.4 It is hard to know how to correct for systematically biased error processes

Suppose I have a compass that points north on average, with some error. Suppose it will be off by at most 90 degrees, and that it is equally likely to be off by $x$ degrees as it as to be off by $-x$ degrees. Knowing this about my compass, I can get a pretty decent idea of which direction north is by simply looking at the compass for long enough. This compass has two nice properties which make this possible: its errors are independent and unbiased, meaning that knowing the size of one error doesn't tell you anything about the size of future errors and that its independent observations average out to a correct answer.

On the other hand, suppose someone took the first compass and observed where it was pointing *once*, and set the compass so that it would tend to point in that direction, again with some random variation around that point. This compass lacks the nice properties of the other compass: its errors are correlated with each other and the compass is biased. The compass can give me *some* indication of which direction north is (it is better than nothing), but I can't use the compass to converge on a correct answer.

The lesson here is that just gathering more data can be very helpful when you have an independent, unbiased error process, but is often less helpful if you have a biased error process or your errors are not independent. When there is more bias or less independence, prior beliefs are more important.

## 2.2 Relevance for moral methodology

Lessons from the Bayesian approach to curve fitting apply to moral philosophy. Our moral intuitions are the data, and there are error processes that make our moral intuitions deviate from the truth. The complete moral theories under consideration are the hypotheses about the phenomena. (Here, I use "theory" broadly to include any complete set of possibilities about the moral truth. My use of the word "theory" does not assume that the truth about morality is simple, systematic, and neat rather than complex, circumstantial, and messy.) If we expect the error processes to be widespread and significant, we must rely on our priors more. If we expect the error processes to be, in addition, biased and correlated, then we will have to rely significantly on our priors even when we have a lot of intuitive data.

### 2.2.1 How the Bayesian approach applies to moral philosophy

The following is an outline of how the Bayesian approach to curve fitting fits in moral philosophy:

| | Science | Moral Philosophy |
|---|---|---|
| Hypotheses about phenomena | Different trajectories of a ball that has been dropped | Moral theories (specific versions of utilitarianism, Kantianism, contractualism, pluralistic deontology, etc.) |
| Hypotheses about error processes | Our position measurements are accurate on average, and are within 1 inch 95% of the time (with normally distributed error) | Different hypotheses about the causes of error in historical cases; cognitive and moral biases; different hypotheses about the biases that cause inconsistent judgments in important philosophical cases |
| Observations | Recorded position of a ball at different times recorded with a certain clock | Intuitions about particular cases or general principles, any other relevant observations |
| Background theory | The ball never bounces higher than the height it started at. The ball always moves along a continuous trajectory. | Meta-ethical or normative background theory (or theories) |

Table 2.1: Curve-fitting in science and moral philosophy

This is not just an analogy. I am claiming that intuitions are a kind of data and that our credences in moral theories should be updated in accordance with our best epistemic theory.

What is included in our priors? The correct, but uninformative, answer to this question is that the prior includes all relevant background information as well as our epistemic standards. Background information could include both information about error processes, our degrees of belief in any relevant meta-ethical claims, and our degrees of belief from prior object-level investigation into normative ethics, or our basic hunches about normative theory. Our basic epistemic standards will include how much weight we give to ideals such as simplicity, explanatory power, generality, coherence, or other theoretical virtues.

What kind of error processes could there be in moral philosophy? It is simplest to understand a scientific error process as some kind of interference with perception (such as an optical illusion), but this is not very similar to moral error processes. The most familiar error process, which I will discuss in Section 2.4, turn on the fact that our intuitive moral judgments are not *direct perceptions*, but the upshot of subconscious processing which uses a variety of shortcuts which lead to errors in certain known circumstances. But given that moral judgments have been subject to widespread

historical error and that even confident judgments by philosophers are subject to error, there are very likely a variety of other bad epistemic influences on our moral judgments.

In this chapter, much of the evidence I present will be about how significant, widespread, biased, and correlated moral error is. I will have less to say about meta-ethics and basic epistemic standards, since these issues are best treated elsewhere. If I'm right, the implication is that we must rely more on our priors when doing normative and applied ethics.

### 2.2.2 Objection: moral philosophy is a priori and requires different methodological standards

Someone may object to my Bayesian approach to moral philosophy in the following way:

> Bayesian approaches to reasoning may be appropriate in scientific contexts, but they are not appropriate for moral philosophy. Moral philosophy is unlike science because it is a priori, and therefore we should expect that different standards of good reasoning apply.

I question the assumption that moral philosophy is a priori, but I will not press the point. The better reply is that the Bayesian approach outlined above is plausible in paradigmatic cases of a priori reasoning. I will argue for this by considering how my argument would apply if we were doing mathematics with limited cognitive ability (so we couldn't always directly prove the correct answer).

Imagine we are testing some generalization, such as "Every differentiable function is continuous," but we don't know how to prove it formally.[4] We have a (perhaps imperfect) method of deciding whether a particular function is continuous and/or differentiable, but we lack the insight to apply it to the general case. We have looked at many cases and every time we've identified a function as differentiable, we've also identified it as continuous. At this point, how confident we should be in the generalization depends a lot on (i) the prior probability of the generalization, which will be influenced by our background information and epistemic standards, and (ii) our information about possible error processes in our methods for determining whether functions are continuous.

If we expect our continuity tests to fail 20% of the time, we should be much less confident in the generalization, and update less as we find more confirming (or disconfirming) instances of the generalization. If we think our errors are independent, we could simply do our continuity tests many times to become very confident that our answers were correct: if we do 100 tests and find that about 80 of them say the function is continuous, we can be very sure that it is continuous. If our errors are not independent, the problem cannot be solved by repeated tests. Suppose our method of testing

[4] For people rusty on calculus, this generalization is true.

for continuity always makes the same error on certain types of problems, but we don't know which types. In this context, we may quickly reach a point where additional tests will not help us, and we'll have to make sure we've used appropriate priors, paid sufficient attention to the error processes, etc. For similar reasons, if we think we found a few functions that we "showed" were differentiable but not continuous, we should place less weight on this finding if our errors are likely to be correlated and if we are more likely to mess up in our proofs.

Now, there are famous problems about developing a theoretical account of Bayesian reasoning that works across mathematical contexts, and I don't intend to solve them.[5] But the most natural approach to this problem is to treat the mathematical claims as independent propositions and treat the outcome of our procedure as our observations. Math is an a priori discipline if anything is, and the Bayesian approach is plausible here. Therefore, the claim that my argument ignores a relevant difference between a priori and a posteriori disciplines is implausible.

### 2.2.3 If we relied on priors more, how could this affect moral philosophy in general?

Suppose we treat as our observations the entire set of intuitions about test cases that moral philosophers have constructed. How might we analyze this data differently if we believed that error processes generating these judgments were more pervasive? What methodological morals could we draw?

#### 2.2.3.1 Focus on the big picture

One clear conclusion is that we should become more suspicious of arguments that reject an otherwise good theory on the basis of a only a few kinds of intuitive counterexamples. If we believe error processes are more significant than we thought, we should be less confident in our intuitive counterexamples, and we should be more willing to write off the counterexamples as errors.

A philosopher may object: a theory can be no more plausible than its least plausible implication. There is a sense in which this is true but doesn't affect my argument, and a sense in which it is false, but would affect my argument if it were true. What's true is that for any two claims $A$ and $B$, if $A$ implies $B$, then $A$ is no more probable than $B$. This is a simple consequence of probability theory. Now, suppose we have some intuitive counterexample $X$ to a moral theory. And suppose that $X$ is the deliverance of some process that is correct 60% of the time. Question: does it follow that we should have less than 40% credence in any theory that implies that $X$ is false?

[5] See Garber (1983) for an introduction to the problem of logical omniscience.

The answer is No. If the information above were our total evidence, then it would be rational for us to have 60% confidence in $X$, and no more than 40% confidence in any theory that implies that $X$ is false. However, suppose that we have a lot of additional data in addition to $X$. If some theory fits all of this data well and there is no alternative theory of comparable scope and prior plausibility that fits all of the data well, then our total evidence might make $X$ much less likely than it otherwise was. In this case, the counterexample may not be very damaging. (If this remains puzzling, ask yourself how you would react to someone who rejected a prediction curve that fit most of the data pretty well, but missed a few points. Do you want to say that the curve is almost definitely wrong on the grounds that it missed five data points that came from a process that delivers correct answers 60% of the time?)

Do philosophers actually reject otherwise good theories on the basis of a few strong counterexamples? It is hard to pin down particular philosophers doing this in print because there are usually many considerations at stake. However, I have seen philosophers reject utilitarianism (about value, not necessarily the consequentialist part) primarily on the basis of the intuition that no number of headaches could be worse than a death, egalitarianism primarily on the basis of the Levelling Down Objection, the Doctrine of Double Effect primarily on the basis of the Loop Case, hedonism (about well-being) primarily on the basis of Nozick's Experience Machine, and subjectivism primarily on the basis of Parfit's Agony Argument (Parfit, 2011). Now, there may be a lot of other good reasons to reject these positions, but none of us should have ever been very convinced by hearing one of these objections all by itself, at least if we ever thought these theories had much going for them in the first place. This would be like rejecting a simple curve that got the broad strokes of the data right, simply on the grounds that the theory missed a few of the data points.

Someone might reply:

> In these cases, there are actually *many* data points that the counterexampled theories are missing. It is easy to construct many variations of the headaches vs. death trade-off, the Levelling Down Objection, the Loop Case, the Experience Machine, and the Agony Argument. Therefore, it is not as if these theories just missed *a few* data points, they are missing entire sections of the curve!

If a hypothesis doesn't match a large section of the data and the error processes along that region of the data are independent, that is strong evidence against the hypothesis. Therefore, this reply would be reasonable if our errors were independent. When our errors have a common cause, as they do in the case of the biased compass, getting additional data that is likely to be subject to the same

kind of error gives us essentially zero new information. However, if our judgments in the cases of the famous counterexamples are affected by an error process, our errors are likely to be correlated with each other in analogous cases. A moral theory with counterintuitive implications about a variety of very analogous cases is like a hypothesis about which direction is north that is violated by many measurements of a biased compass with correlated error. Many errors about analogous cases are not much worse for a theory than counterintuitive implications about a few such cases. Hence, this objection fails.

In short, the conclusion here is that *if* our judgments are subject to significant, widespread, correlated error processes, as I will be arguing in later sections, then we can draw the following moral:

> A few counterexamples are not strong evidence against a moral theory that does reasonably well with respect to most of the intuitive data, one's background theories, and basic epistemic standards (though they are some evidence against them). Evidence is much stronger if we can provide multiple kinds of counterexamples.

How can this moral be applied?

Moral philosophy could be improved if philosophers would consider a wide variety of cases as their data, note which theories have implausible implications about which cases, and evaluated the result in a holistic way, maintaining a willingness to accept theories that didn't satisfy all of their intuitions. Directly applying any Bayesian approach is not a practical option yet, but we can internalize reasonable epistemic standards, look out for errors, and informally try to determine which of the competing theories does best in terms of fit with data and prior probability.

Before concluding this subsection, I should acknowledge an important qualification to my claim that a few different kinds of counterexamples are not strong evidence against a moral theory that does reasonably well with most of the intuitive data, one's background theories, and basic epistemic standards.

There are important distinctions between different circumstances under which a counterexample is leveled against a theory. Some counterexamples arise because a theory has very different implications than the authors of the theory intended it to have, some counterexamples are very different from the types of cases that had been considered when constructing the theory, and some counterexamples rely on a very ordinary case about which people have firm opinions or a case that is similar to a very similar to such a case. If a counterexample has all of these properties, it is often enough to sink a theory that functions very well in other circumstances. Consider, for instance,

Parfit's Hell One and Hell Two counterexample to average utilitarianism (Parfit, 1984, 393). In Hell One, some number of people exist and suffer greatly. Hell Two is just like Hell One, but there are many additional people who exist and suffer slightly less. Average utilitarianism implies, absurdly, that Hell Two is better than Hell One because the average level of suffering is lower. If someone had not been very careful, they might be attracted to average utilitarianism because they think it formalizes the intuition that "quality is more important than quantity." But when people had this intuition, they were mainly thinking about cases involving trade-offs between quality and quantity for people that are well off rather than badly off. Because of this, the counterexample has the first two features that I said make a counterexample stronger. It also has the third because, in general, people are confident that, other things being equal, it is bad if additional people have bad lives. Because the counterexample has these properties, it seems that if we were originally attracted to average utilitarianism in these circumstances, it is because we hadn't considered certain types of strong evidence and we didn't really understand what we were signing up for. Appreciating this kind of counterexample would give us very strong reasons to reject average utilitarianism in these circumstances.

#### 2.2.3.2 Give more weight to simplicity, or whatever your basic epistemic standards endorse, than you otherwise would

Those who accept my argument might more often appeal to basic epistemic standards, such as simplicity, in order to decide between two theories differently than they otherwise would.[6]

To illustrate this idea, consider what would happen if our basic epistemic standards favor moral theories that are simple. I believe this might tip the scales for individuals in the following sorts of epistemic states:

1. Consider someone conflicted between accepting a simple Doctrine of Double Effect analysis of constraints and a more complicated Kamm-style view that had better fit with intuition. This person might update significantly in favor of the Doctrine of Double Effect.

2. Consider someone whose credence is split between rule consequentialism and a form of pluralistic deontology. Presumably, the pluralistic deontology has better fit with intuition, but will be more complicated because it has a larger number of adjustable parameters. This person might update significantly in favor of rule consequentialism.

[6]Some moral philosophers, such as Nagel and Temkin, have expressed deep skepticism about the value of simplicity as a basic epistemic standard for moral philosophy (Nagel, 1991, x-xii), (Temkin, 2012, 18-19). I will not attempt to address this issue here, since my arguments will not make heavy weather of simplicity as a basic epistemic standard.

3. Finally, if someone read this chapter and came to believe that moral error was much more extensive and correlated than they previously expected, and they had not realized the methodological implications of this, they might update significantly in favor of a view like act consequentialism. This might make sense because they would be much more moved by act consequentialism's simplicity and elegance, and much less moved by its departures from intuition.

Obviously, I am not going to succeed in proving that updates of any particular size will be justified with any serious degree of precision. This would require a lot of steps, such as assigning prior probabilities to the hypotheses in question, identifying data to update on, specifying the error processes more precisely, and performing calculations. But I do have a strong case that those who already accept simplicity as a theoretical desideratum should find these simpler theories more plausible than they otherwise would have. How much more plausible they should find them will call for personal judgment. The next three sections help clarify our picture of the error processes that contribute to moral judgment, so that we can have a better idea about how much less to trust intuition and how much more to rely on our priors.

## 2.3 The historical record

This section examines the historical record of moral error and addresses its implications for what kind of error processes we should expect to affect our moral judgment. I will be brief in stating my argument:

1. In the past, there was widespread, correlated, biased moral error, even on matters where people were very confident.

2. By induction, we make similar errors, even on matters where we are very confident.

The first premise is very plausible to anyone who reflects on the history (and current distribution) of moral opinion on the treatment of children, women, slaves, racial and ethnic minority groups, gays and lesbians, animals, the poor, immigrants, prisoners, religious and political dissidents, drug users, or just about any group that has ever been marginalized in a serious way. These errors were widespread, in that for any particular person subject to them, they affected a significant share of the person's moral reasoning. They were correlated, in that there was a pattern to the bias which could not be avoided by simply considering variations of similar cases, and the errors were biased in that they did not average out to reasonable views. (In speaking of these examples in a historical context, I don't mean to suggest that people no longer make these errors.)

There is something helpful about the brief remarks above, but the first premise has more independent plausibility than anything I can say to support it, so I will stop here and consider objections to the argument. The first category of objections denies the induction, appealing to differences between the past and the present. The second category tries to downplay the significance of previous error.

### 2.3.1 Against the induction step

**Objection: Past errors were largely self-serving parochial prejudices, rather than theoretical errors about things like population ethics or the philosophy of risk. Even if we should expect more errors driven by self-serving parochial prejudices, why think that such errors should tell us anything about the importance of shaping the far future?**

Neglect of future generations is largely caused by lack of neutrality between present and future generations, and intrinsic preference for benefits to come sooner rather than later. We can argue about the defensibility of these preferences (and I will do so at length in later chapters), but they clearly serve the interests of a small class of people (presently existing people) who have power over a large class of other people (future people) who cannot defend their interests. If these preferences are morally wrong, they fit the pattern of self-serving parochial prejudices remarkably well.

**Objection: Past errors were mostly due to limited and/or inaccurate non-normative information. Since we have much more information, we should expect much less moral error.**

I agree that we should expect less moral error now than in the past, but I suspect that we will continue to be subject to very significant amounts of moral error. I have three replies that illustrate why. The replies are speculative, but so is the objection.

First, (Haidt, 2001) finds that people tend to arrive at their moral judgments by unconscious emotional processes, and then come up with post hoc rationalizations which they use to defend their judgments to others. When old excuses, like "animals don't feel pain" get disproven, people can always come up with new ones. I therefore suspect, but cannot prove, that a large fraction of moral error would remain, with merely different post hoc verbal justifications, if people had more information.

Second, there is little reason to believe we have uncovered all, or even most, of the error-relevant information we might be lacking. It would be surprising if we were the first generation to be mostly

free of these mistakes.

Third, even if it were true that moral errors were mostly caused by lack of non-normative information and we had uncovered all of the information, we still may not be out of the woods. If it can help us avoid moral error, information must be appropriately internalized, rather than simply available (on Google, say), which presents an additional challenge. I would bet that even if new information were available, politically convenient errors would persist despite readily available contradictory evidence. For example, consider the fact that typical Americans believe that about 20% of the US government budget is spent on foreign aid, and wish it only spent about 10%. (In fact, the US government spends less than 1% on foreign aid.)

### 2.3.2 Against the claims about the historical error processes

**Objection: On some meta-ethical views, historical moral error was probably much less abundant than on others. For example, if morality is really just about idealized preferences, it is unclear that people from the past made significant moral errors relative to their own standards. It is therefore unclear that we will make moral errors relative to our own standards.**

I agree that if morality is really just about idealized preferences, then error from the past may be less than it may seem at first glance. However, it is independently plausible that there have been significant errors in the past, and if a meta-ethical view cannot make sense of that, then this is a cost of adopting that meta-ethical view. Rather than appealing to idealized preference views to support the claim that there have not been significant historical errors in moral reasoning, it would seem more appropriate to develop a version of this view that allowed us to accommodate this datum.

**Objection: We have reason to believe that philosophers would not be subject to whatever these historical error processes were.**

While we may reasonably expect that philosophers would be somewhat less error prone than the general public (after all, they expect their intuitions and verbal justifications to be subject to unusual levels of scrutiny), the historical track record does not inspire a great deal of optimism. It isn't hard to find moral and political philosophers of the highest caliber expressing outrageous moral positions. Aristotle endorsed slavery; Heidegger was a Nazi; Hume and Marx have their racist moments; and Kant offered a variety of contorted justifications for atrocious opinions about women, homosexuality, masturbation, children born out of wedlock, organ donation, and rebellion against unjust states

(Schwitzgebel, 2010). In fairness, these philosophers may have had more plausible views than many of their contemporaries, and others were able to see through significant errors of their day (consider, for example, Mill's enlightened views on animal welfare, the subjugation of women, slavery, and the treatment of gay people). Nonetheless, it is not plausible that philosophers are largely immune to normal human error.

## 2.4 The scientific record: biases

The main claim of this section is the following:

1. Scientific findings on prudential, epistemic, and moral heuristics and biases strongly suggest that our moral judgments are subject to error processes which are widespread and biased.

2. Because of this, we should expect philosophers' intuitions to be subject to error processes which are widespread and biased.

I will focus on some biases that are relevant to evaluating the importance of shaping the far future, but also aim to give an overview of the kinds of biases have been discovered already.

### 2.4.1 Background on heuristics and biases

The research on heuristics and biases largely stems from seminal work by Kahneman and Tversky during the 1970s.[7] They found that people use many *heuristics*—shortcuts essentially—in their reasoning. These heuristics are a fairly reasonable way to handle complex information and make decisions, but they lead to certain kinds of predictable errors—errors that don't, on average, get a correct answer—known as *biases*. To claim that we are subject to some bias is partly to make a normative claim. It involves claiming both that we do reason in some way, and that, at least in certain circumstances, we ought not to reason in that way. In most of the heuristics and biases literature, the patterns of reasoning are obviously bad, so the normative work is not too hard, and people sometimes forget that this is not a purely scientific enterprise. An enormous number biases have since been discovered since Kahneman and Tversky started their work. At the time of this writing, Wikipedia lists 93 biases in the category "decision-making, belief, and behavioral biases."

[7] For an overview of the field, seeDaniel Kahneman and Tversky (1982), or, more recently, Kahneman (2011).

### 2.4.2 Prudential biases

By *prudential* biases, I mean biases where people make errors in pursuit of their self-interested goals. These biases are fairly easy to discover, since we can test whether people who make these errors will tend to do worse at achieving their self-interested goals.

An example of such a bias is what Parfit called the *bias toward the near*, or what some social scientists would call *hyperbolic discounting*. When this bias is present, people prefer benefits to come sooner rather than later, but their preferences are inconsistent over time. For example, while many people would prefer to get $10 today instead of $15 next month, few would prefer to get $10 in 12 months instead of $15 in thirteen months.[8]

A second example of such a bias is *probability neglect*. When probability of some risk or payoff is small, willingness to pay to avoid the risk or get a chance at the payoff is not sufficiently sensitive to the probability. To get a flavor for this bias, note that as long as they aren't thinking about both options at the same time, people's willingness to pay for insurance against risks of 1/100,000, 1/1,000,000, and 1/10,000,000 are essentially equal, though they should each differ by a factor of roughly 10. Similar results were found for 1/650, 1/6300, and 1/68,000.[9] The problem does not totally go away when people think about both options at the same time. Willingness to pay to reduce the risk of a serious injury by 12/100,000 was only 20% higher than willingness to pay to reduce the same risk by 4/100,000 (Jones-Lee et al., 1995), though expected utility theory (very plausibly) mandates that people should be willing to pay about three times more.

### 2.4.3 Epistemic biases

By *epistemic* biases, I mean common patterns of reasoning which systematically lead people toward unreasonable probability assignments. These biases are readily identified by showing that certain patterns of reasoning often lead to verifiably wrong answers.

For example, people have a truly remarkable bias toward *overconfidence*. If people were well-calibrated, then, in general, when they were 90% confident in some proposition, they'd be correct about 90% of the time. Sadly, almost everyone is overconfident. To test this, researchers have asked people for their confidence intervals at different levels of probability for a wide variety of ordinary questions. To say that your 98% confidence interval for the height of Mt. Everest is 27,000 ft to 33,000 ft is to say that you are 98% confident that the height is somewhere in that range, and that you think it is equally likely to be above or below it. One study found that when people gave their

[8] See (Thaler, 1981) for examples of this sort.
[9] See (Sunstein, 2002) and (Kunreuther et al., 2001) for further details.

98% confidence intervals for a wide variety of different questions, the true value was outside these intervals about 46% of the time. This effect does not go away when you warn people about it (Alpert and Raiffa, 1982).

Another significant epistemic bias is *hindsight bias*, also known as the "I knew it all along" effect. I quote from another paper summarizing another studies findings related to this effect:

> (Fischhoff and Beyth, 1975) presented students with historical accounts of unfamiliar incidents, such as a conflict between the Gurkhas and the British in 1814. Given the account as background knowledge, five groups of students were asked what they would have predicted as the *probability* for each of four outcomes: British victory, Gurkha victory, stalemate with a peace settlement, or stalemate with no peace settlement. Four experimental groups were respectively told that these four outcomes were the historical outcome. The fifth, control group was not told any historical outcome. In every case, a group told an outcome assigned substantially higher probability to that outcome, than did any other group or the control group. (Yudkowsky, 2008)

### 2.4.4 Moral biases

By *moral biases*, I mean patterns of reasoning which lead people into morally defective actions, preferences, or beliefs. Moral biases are harder to detect than other biases because we do not have readily agreed upon standards for moral error. It is much easier to establish that a certain kind of strategy fails to achieve the goals of the person pursuing it, or that a certain pattern of reasoning tends to lead to incorrect beliefs about readily verifiable propositions. But there are some clear, and disturbing, examples of moral bias.

One such bias is *scope insensitivity*. The finding is that when the numbers are large, but even sometimes when they aren't, people are insensitive to the difference between harms/benefits with 10 or even 100 times greater scope. A harm or benefit has greater *scope* when it affects more people, more animals, a greater portion of the environment, etc. For example, as long as they aren't thinking about both options at the same time, people are willing to pay essentially nothing more to prevent 9,000 people from passing a kidney stone than they are to prevent 90,000 people from passing a kidney stone (Baron and Greene, 1996). There is nothing special about kidney stones, people react similarly to number of birds saved from drowning in oil spills, number of lives saved, and many other types of harms. In these cases, people make extreme errors, since they evaluate options that are 10 or 100 times different as if they were equally good. This behavior also cannot be explained by

budget constraints.

People also engage in illicit forms of *proportional reasoning*. Fetherstonhaugh et al. (1997) found that participants significantly preferred helping a fixed number of people in a refugee camp when the *proportion* of people helped was greater. Describing the participants' hypothetical choice, they write:

> There were two Rwandan refugee programs, each proposing to provide enough clean water to save the lives of 4,500 refugees suffering from cholera in neighboring Zaire. The Rwandan programs differed only in the size of the refugee camps where the water would be distributed; one program proposed to offer water to a camp of 250,000 refugees and the other proposed to offer it to a camp of 11,000.

Participants significantly preferred the second program. In another study, Slovic (2007) found that people were willing to pay significantly more for a program of the second kind.

I regard these preferences as relatively clear errors, but others may try to justify this reasoning. I am skeptical of this move for several reasons. First, this seems like a clear instance of a more systematic bias which affects people in prudential contexts as well. For example, people are willing to walk a few blocks to another store to save $50 on a $150 jacket, but unwilling to walk a few blocks to another store to save $50 on a $2000 cruise ticket. Second, the fact that one program would save *all* the lives of people in some socially salient grouping does not seem like a strong reason to save their lives, rather than the lives of twice as many people that do not constitute the entirety of a socially salient grouping.[10] If this were not true, then Oxfam could increase the strength of my moral obligations to aid just by breaking up a larger refugee camp into smaller socially salient groups.[11] Since I do not believe that this would make my moral obligations stronger, I find it much more plausible to believe that I have been deceived by a tendency to look at proportions when absolute values are what matters.

A third moral bias, probably related to the first two, is the *identifiable victim bias*. In general, people are much more motivated to help an identifiable victim than they are to help "statistical" victims or unidentifiable victims. Here are two examples. First, people are more willing to help 1 identified child (say, with a name and a picture) than to help 8 unidentified children (Kogut and Ritov, 2005). Second, Singer (2009, 48) reports on a study by Small and Loewenstein (2003):

[10] Unger (1996) expresses a similar view.

[11] Someone may suggest that morality should be set up so that no agent has the power to "manipulate" my moral obligations in this way, but proportions really do matter. This theory seems to, absurdly, predict that if a large storm divided a refugee camp into socially salient groups, then it would increase my moral obligations to aid.

> People asked by researchers to make a donation to Habitat for Humanity in order to house a needy family were told either that the family "has been selected" or "will be selected." In every other detail, the wording of the request was the same. . . .Yet the group told that the family had already been selected gave substantially more.

It's possible that there is some reason to favor identifiable beneficiaries *slightly* because if the beneficiary is known in advance, it might make it slightly more likely that a real person ever gets helped. After all, finding a reasonable beneficiary is one step that could go wrong. But this hardly seems like enough to justify the observed differences in our judgments.

### 2.4.5 We should expect that there are many unknown moral biases

Because moral biases are hard to detect, but we know that they exist, we should expect that there are many moral biases which we have not yet detected. To appreciate this on an intuitive level, consider a case in which a child drops some BBs on the carpet. Imagine that we know that the child dropped some large BBs, which are much easier to see, and some small BBs, which are harder to see. Even if we have found more large BBs than small BBs, we may rationally expect that there are many more small BBs in the carpet, since we are able to detect a smaller proportion of the total amount of small BBs.

So it is with prudential/epistemic biases and moral biases. We have found more prudential/epistemic biases than moral biases, but moral biases are much harder to detect. Therefore, we should expect to have discovered a greater proportion of the prudential/epistemic biases than the moral biases. So we should expect that there are more moral biases than the ones we have discovered (an even greater proportion than the proportion of undiscovered epistemic and prudential biases, the known number of which has been steadily growing over time).

### 2.4.6 Philosophers are probably subject to moral bias, just like everyone else

I have now argued that we are subject to a variety of biases that impair prudential, epistemic, and moral reasoning. Absent some special reason to the contrary, we should expect philosophers to be subject to these error processes as well. Is there some reason to resist these arguments?

One might hope that now that we know about these biases, we can simply look out for them and avoid them, but otherwise carry on with our reflective equilibrium as usual. This hope would be misplaced for at least two reasons. First, as argued in the previous section, there are likely to be

many unknown moral biases. Not knowing what biases to expect, we cannot simply identify them and try to avoid them. Second, knowing about our biases doesn't make it easy to overcome them. People who know about a bias still have the bias, they just *think* that they suffer from it less than average (Pronin and Kugler, 2007). For example, people still think they're better than average, even when you warn them about the better than average bias. Likewise, people still exhibit hindsight bias and overconfidence after being specifically warned about those biases.

Alternatively, one might hope that philosophers would be more reflective and rational than ordinary people. One might hope that this would allow them to avoid the more destructive effects of moral bias. There is limited data available on this issue, but the findings we have so far support the hypothesis that, more or less, philosophers suffer from the same biases as everyone else. In a study by Schwtizgebel and Cushman, philosophers' intuitions exhibited order effects while considering test cases for the doctrine of double effect, the action-omission distinction, and the principle of moral luck. In other words, as with ordinary people, their moral intuitions depended on the order in which the cases were presented (Schwitzgebel and Cushman, 2012).

One conclusion we can draw is that when some argument depends crucially on an intuitive judgment about a particular case, philosophers would do well to examine whether this judgment is likely to be influenced by any known biases. If it it fits the pattern of a known bias, then that can dramatically increase the probability that the judgment is merely the product of a bias, thereby significantly weakening the argument. I predict that, if challenged on a particular concrete case, philosophers will find themselves confident that their judgments are not the product of a mere bias. Given what we know about the bias blind spot, such protests should be met with a dose of skepticism.

## 2.5 The philosophical record

The main argument of this section is as follows:

1. A number of impossibility results show that certain moral judgments about which philosophers are very confident are inconsistent with each other. And this happens especially often in population ethics.

2. Therefore, we should expect that there are some error processes underlying these judgments that biases us toward overconfidence, and we should not expect to find a theory of population ethics which accords with all of our most confident moral judgments.

3. A relatively limited amount of resources (the careers of a few very insightful philosophers)

generated most of these impossibility results.

4. This search process is unlikely to have uncovered a significant proportion of the important impossibility results.

5. Therefore, we should expect that there are many more such impossibility results.

6. Therefore, we should expect that analogous error processes are operating in many cases where impossibility results have not yet been discovered, and that it will be impossible to find theories that accord with all of our most confident moral judgments in these cases either.

The argument mostly speaks for itself, but I will spend some time justifying the claim about the existence of these impossibility results, and the process that generated our knowledge of these results.

### 2.5.1 Impossibility results

It would take too long to present many of the troubling impossibility results referred to above, so I will briefly discuss a few of them. The existence of these results establishes the first premise of the argument above.

#### 2.5.1.1 Parfit and Arrhenius on population ethics

The first such result is Parfit's Mere Addition Paradox (Parfit, 1984, ch. 19), which I will present *very briefly*. Consider the diagram below:

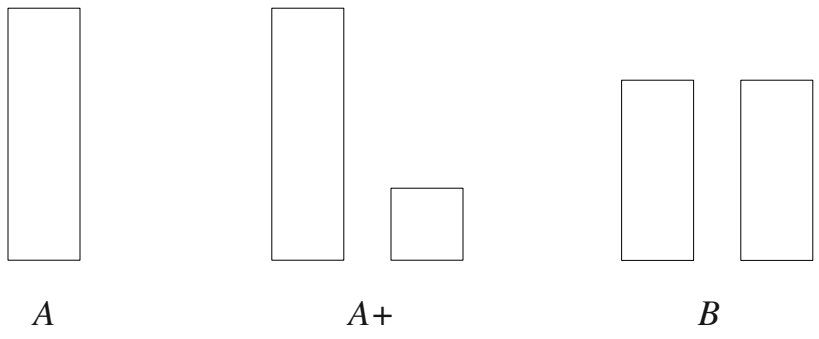


Figure 2.3: The Mere Addition Paradox

The height of a column represents the quality of life of a people in a group. The width represents the size of the population. The first alternative, $A$, represents a case where 10 billion people exist with a very high quality of life. In the second alternative $A+$, the same people exist with an equally high quality of life. But there is also a totally separate population of 10 billion people with a lower, but still very high, quality of life. In $B$, the people in the first group are significantly worse off than in $A$ or $A+$. However, the second group of people is very much better off than they were in $A+$. These effects together make each group equally well off, though they are much worse off than any of the people in $A$.

The paradox is that each of the following judgments are compelling, but they could not all be true:

1. $A+$ is not worse than $A$
2. $B$ is better than $A+$
3. $A$ is better than $B$
4. Transitivity: for any alternatives $X$, $Y$, and $Z$, if $X$ is better than $Y$ and $Y$ is better than $Z$, then $X$ is better than $Z$.

It is intuitive that $A+$ is not worse than $A$ because it is hard to believe that merely creating people with lives worth living could make things *worse*, especially when the extra people have a high quality of life. And it seems that the inequality could not be bad, provided it arose in this way. It is intuitive that $B$ is better than $A+$ because there is a greater total amount of well-being and less inequality. And it can plausibly be argued that $B$ is worse than $A$ because it seems that when there are "enough" happy people, quality of life matters more than quantity of life.[12]

This is one example of an impossibility result in population ethics. In his book manuscript, Gustaf Arrhenius has identified *five* additional impossibility results in population ethics, most of which I would say are more surprising than the Mere Addition Paradox (Arrhenius, 2013).

#### 2.5.1.2 Temkin's Spectrum Paradoxes

Larry Temkin's Spectrum Paradoxes are another class of troubling impossibility results. We will discuss these paradoxes at greater length in other chapters. It is best to start with an example.

[12] I confess that while I find this last judgment intuitive, I don't find it deeply compelling. It is not the kind of thing I would be surprised to find that my intuition were wrong about. However, I find the judgment more plausible when the size of $A$ is much larger, say $10^{80}$ instead of 10 billion. All the same, I have included the example because many philosophers find each premise of this impossibility result very hard to deny.

Consider the following spectrum of alternatives, described in the abstract. We'll call it the "Shock Spectrum."

($A_1$) $2^1$ people are given non-fatal shock treatment for $10(.9)^1 = 9$ years

($A_2$) $2^2 = 4$ people are given non-fatal shock treatment for $10(.9)^2 = 8.1$ years

($A_3$) $2^3 = 8$ people are given non-fatal shock treatment for $10(.9)^3 \approx 7.29$ years

$\vdots$

($A_{200}$) $2^{200} \approx 1.61 \times 10^{60}$ people are given non-fatal shock treatment for $10(.9)^{200}$ years $\approx$ 0.22 seconds

As (Temkin, 1996, 2012) has pointed out, this kind of spectrum presents a puzzle, at least if you play around with the numbers and make the case appropriately boring. The puzzle is that all of the following things are plausible, but couldn't all be true:

1. For each $i$, since $A_{i+1}$ involves twice as many people suffering for 90% as long as $A_i$, $A_{i+1}$ is worse than $A_i$.

2. Since $A_{200}$ involves such a small burden (per person) and the two people in $A_1$ suffer so badly, $A_1$ is worse than $A_{200}$.

3. Transitivity: for any alternatives $X$, $Y$, and $Z$, if $X$ is better than $Y$ and $Y$ is better than $Z$, then $X$ is better than $Z$.

This Spectrum Paradox trades off size of personal burdens (including intensity and duration) against the number of people bearing the burdens. The essence of the paradoxes in this class is that in many circumstances, we must trade off different "dimensions" of goodness against each other, such as quality of a good (per unit time), duration, number (of people enjoying the good), and probability (of enjoying the good). We are typically willing to sacrifice a small amount of one dimension in order to get a large amount of the other dimension. But small costs can add up to large costs, and we may be unwilling to make a very large sacrifice along one dimension to achieve a great gain along another dimension. Such paradoxes can be constructed using any two of the "dimensions" mentioned above, and along multiple dimensions. These paradoxes are deeply troubling, since they contain premises which all seem obviously true, but some of the premises must be false.

### 2.5.2 What to make of the impossibility results

When we learn about these impossibility results, it is rational to temper our confidence in the judgments that are part of the impossibility results. If we originally had very high probability in all of the inconsistent premises of the impossibility result, we must decrease our confidence in some of those judgments.

Someone may believe that these impossibility results will cause problems only within circumscribed domains. I would not be optimistic. As with our discussion of moral biases (and our discussion of the BBs), we must consider the process that drew these results to our attention. These impossibility results were generated from the careers of a few ingenious philosophers who went looking for them. It would be surprising if we had found all of the results already.

We know simply from looking at the impossibility results discovered so far that we will not find a moral theory that fits with all of the moral judgments that people regard as obviously correct. The most we can hope for is damage control. Our reflective equilibrium could be, at best, a system that preserves most of our values.

Once we know about these impossibility results, we may also become more sympathetic to developing complete moral theories, rather than piecemeal ones. One thing that militates against developing complete theories is the desire not to create a theory which is subject to intuitively devastating counterexamples. If one creates a theory the covers a small corner of the moral universe, one can be relatively safe from such counterexamples. In contrast, it will be relatively easy to find counterexamples to a complete theory with clear implications about everything (indeed, we can just go look at our list of impossibility theorems and see which conditions are violated!). But given that we can do this, we know already that the piecemeal theories will suffer from some of these problems, however they are developed. So we should not penalize the clear and systematic theories for having implausible implications in these areas. And we should keep in mind what we know about the process that generates our evidence: a vague and indeterminate theory may be hard to apply, and therefore hard to counterexample, but that shouldn't make us think that there are fewer good counterexamples out there.

## 2.6 Special relevance for the value of shaping the far future

Consider again the conclusion of the rough future-shaping argument:

> From a global perspective, what matters most (in expectation) is that we do what is best

(in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

This claim is surprising, though not crazy.

What is more surprising is the claim that even very small effects on the far future—such as reducing existential risk by 1 in a million—are overwhelmingly important. Some might be tempted to reject my thesis immediately on the grounds that it does not accord with common sense. We should resist this response for two reasons. First, as I will later argue, this conclusion is supported by some generalizations with great plausibility. When a generalization has independent support, we should be especially suspicious that isolated counterexamples to the generalization are generated by an error process. Second, given the complexity and unfamiliarity surrounding the evaluation of shaping the far future, directly intuiting its value seems unreliable. Third, taking account of many specific error processes discussed in this chapter would lead one to distrust direct intuitions about the importance of shaping the far future. I will be pressing the second and third points in this section.

### 2.6.1 More general reasons to distrust direct intuition about the value of shaping the far future

In 2010, Jeff McMahan wrote a *New York Times* opinion piece where he defended the following thesis: if it were possible to eliminate predators without negatively affecting ecosystems in a significant way, it would be desirable to do so. His reason for this was that predators cause animal suffering, and animal suffering is bad, even when it is "natural" (McMahan, 2010). Some philosophers commenting on the story suggested that the implication that there could be a weighty reason to prevent wild animal suffering was so absurd that it was a reductio of the "utilitarian" basis on which McMahan defended his conclusion.[13] McMahan didn't defend his view by appealing to utilitarianism, but that's not what's interesting about this story. What's interesting is that some philosophers would think that this would be a good reason for rejecting a theory like utilitarianism.

It seems that this was a case where intuition was not to be trusted: people don't know very much about wild animal suffering; the context of action is complex, unfamiliar, and vast; the consequences of different interventions could be enormous; and the basic philosophical issues of animal welfare are thorny. Empirical moral psychologists tell us that intuitive moral judgments are the product of fast and frugal heuristics,[14] and these heuristics seem especially likely to fail in a context like this.

[13] See the comments on Erler (2010).
[14] See (Haidt, 2001) and (Greene and Haidt, 2002) for support of this claim.

For very similar reasons, I would have a similar reaction to someone who thought it was absurd that shaping the far future should be such a major altruistic priority: people don't know very much about the far future; the context of action is complex, unfamiliar, and vast; the consequences different interventions could be vast; and the basic philosophical issues about future people, uncertain risk, and small probabilities are thorny. For quite general background reasons, direct appeals to intuition about the value of shaping the far future seem unlikely to get us very far.

### 2.6.2 How the biases discussed above affect judgments about the value of shaping the far future

When we zoom in on the biases we discussed in Section 2.4, we find additional reason to distrust direct intuitions about the value of shaping the far future.

Given scope insensitivity, we should expect to fail to fully appreciate the value of many future generations. It is unlikely that our intuitions are appropriately sensitive to the order of magnitude of how long earth-originating intelligent life survives and thrives. Since much of the expected value of shaping the far future comes from the scenarios where we survive and thrive for an exceptionally long time, we should expect that people suffering from scope insensitivity would underestimate the importance of this factor.

Since no future lives are identifiable, nearly all the benefits of shaping the far future are essentially statistical. The identifiable victim bias suggests that we would therefore underestimate the value of shaping the far future.

On the basis of probability neglect, we should expect to be insufficiently sensitive to small changes in probability of an outcome, especially when the outcome already has a small probability. This would again make us underestimate the value of reducing existential risk or producing other positive trajectory changes, since the probability that any action significantly changes the course of history is likely to be small.

Since the benefits of shaping the far future are distant in time, given our bias toward the near, we should expect to underestimate the value of shaping the far future.

Most people seem to believe that the probability of an existential catastrophe this century is very low. Given what we know about the bias toward overconfidence, we should expect people to be overconfident about this, and we should expect this to negatively affect their intuitive assessment of the value of efforts aimed at reducing existential risk. For a mostly non-overlapping discussion of cognitive (contrasted with practical) biases affecting the assessment of global risks see (Yudkowsky,

2008).

Scope insensitivity, identifiable victim bias, and probability neglect tend to be less significant when we quantify impact, probability, and expected value of each alternative, and when we consider the alternatives all at once. I will be doing this as I present my argument in later chapters, so we should expect to find it increasingly plausible that it is overwhelmingly important to shape the far future.

## 2.7 Conclusion

Learning about moral errors through history, biased heuristics generating our moral judgments, and a collection of impossibility results should tell us that our moral judgments are subject to errors that are hard to detect and hard to correct. In light of this, we should trust intuition less and rely on our priors more. We should not expect to find a theory that fits all of most confident moral judgments, and we should largely be engaged in an exercise in damage control, especially in population ethics. Finally, we should expect these error processes to lead us to significantly underestimate the importance of shaping the far future.

# Chapter 3

# The Case for Shaping the Far Future

## Introduction

The purpose of this chapter is to clarify and better defend the rough future-shaping argument that I outlined in the first chapter. Recall that the argument goes as follows:

1. Humanity may survive for millions, billions, or trillions of years.
2. If humanity may survive may survive for millions, billions, or trillions of years, then the expected value of the future is astronomically great.
3. Some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways.
4. If the expected value of the future is astronomically great and some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways, then from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.
5. Therefore, from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

Two sections involve very broad speculation about what the far future will be like and how we could shape it. The other section is about the normative assumptions that this argument relies on. The rest of the dissertation will be devoted to analyzing these normative assumptions, and will contain less broad speculation.

In section 1, I defend the first premise of the rough future-shaping argument. In brief, I argue that there is a non-negligible probability that our descendants will survive on Earth as long as the Earth is habitable, and that there is a non-negligible probability that our descendants will colonize space. I don't think anyone knows what the objective chance of these things happening is, but I argue that the subjective probability of these things, in the sense of reasonable betting odds, is neither very low nor very high. If accepted, my argument shows that the expected duration which our descendants will survive is very great, on the order of billions of years or more.

In section 2, I defend four normative assumptions which I will use to support the second premise. Very roughly, the strategy is to estimate the value of the future as a whole by (i) carving up the history the world into large chunks of time, which I call "periods," (ii) assigning a value to each period which says how well it goes, and (iii) adding up the value of all those periods. This is only an approximation, and I don't argue that it would work in all cases, but I argue that it is reasonable in the cases of interest. In addition, some of these assumptions are very debatable, and I analyze their plausibility in much greater depth in later chapters. Readers looking for detailed arguments and responses to objections to these normative assumptions should look at later chapters. The goal in this chapter is to illustrate the plausibility of these assumptions and show how they work together to support my main conclusion.

In section 3, I defend the third and fourth premises of the argument. Part of this is a fairly straightforward application of the results of the previous two sections. In a less straightforward part that involves more judgment calls, I use the results of the previous two sections to informally compare the value of proximate benefits, benefits from speeding up development, and benefits from trajectory changes. I argue that creating positive trajectory changes is the best way to shape the far future. As in chapter 1, I don't take a stand on whether the best ways of creating positive trajectory changes are very targeted or very broad, and I give some reasons to think that very broad candidates may be plausible. At this stage, the most we can say is that (i) it would be very valuable to understand how broad or targeted the best ways of producing positive trajectory changes are, and (ii) many ordinary actions predictably differ significantly in terms of how much they affect the far future, and the most promising broad approaches to creating positive trajectory changes probably pay some attention to this.

## 3.1 How long could humanity survive?

### 3.1.1 How long could life on Earth last?

Consider the following passage from the end of *Reasons and Persons*. Derek Parfit writes:

> I believe that if we destroy mankind, as we now could, this outcome would be much worse than most people think. Compare three outcomes:
>
> (1) Peace.
>
> (2) A nuclear war that kills 99% of the world's existing population.
>
> (3) A nuclear war that kills 100%.
>
> (2) would be worse than (1), and (3) would be worse than (2). Which is the greater of these two differences? Most people believe that the greater difference is between (1) and (2). I believe that the difference between (2) and (3) is *very much* greater.
>
> ... The Earth will remain inhabitable for at least another billion years. Civilization began only a few thousand years ago. If we do not destroy mankind, these few thousand years may be only a tiny fraction of the whole of civilized human history. The difference between (2) and (3) may thus be the difference between this tiny fraction and all of the rest of this history. If we compare this possible history to a day, what has occurred so far is only a fraction of a second. (Parfit, 1984)

What are the chances that we will last this long?

When I say "we," "civilization," or "humanity," I am not just talking about the species *homo sapiens*. Very few species last for anything like 1 billion years, most dying off within 10 million years. I am asking, "How long will people exist in the future?" Here, I mean "people" in the sense of "sentient beings that matter," which, I am assuming, will include our intelligent descendants.

When we include our intelligent descendants, it is not absurd to consider the possibility that civilization continues for a billion years, until the Earth becomes uninhabitable. We can't know the frequency with which civilizations like ours survive that long, or the "objective chance" that we'll survive that long. But we can say something about what reasonable betting odds would be with respect to the claim that our descendants will survive for at least 1 billion years. How likely humans are to survive into the far future depends on what decisions society makes. This may be the most uncertain aspect of our estimate. If people do little to address existential risks, we may be unlikely to survive in the far future. If people cooperate and exercise reasonable caution, we may be very likely to have a long and prosperous future. Given the great uncertainty involved, including

uncertainty about what people will do to prepare for these risks, it would seem overconfident to have a very high probability or a very low probability that humans will survive for the full billion years. Having a very high or low probability in this claim, such as less than 1% or greater than 99%, would require much greater certainty about the future than it is reasonable to have. Obviously, choosing any specific number here would be arbitrary. To be conservative, I will assume that our subjective probability in this claim should be at least 1%. My argument would, of course, only work better if, as I believe, we are more likely to survive this long, since that would increase the expected value of preventing premature extinction or otherwise shaping the far future.

### 3.1.2 Beyond a billion years?

The lion's share of the expected duration of our existence comes from the possibility that our descendants colonize planets outside our solar system. There are many stars that we may be able to reach with future technology (about $10^{13}$ in our supercluster). Some of them will probably have planets that are hospitable to life, perhaps many of these planets could be made hospitable with appropriate technological developments. Some of these are near stars that will burn for much longer than our sun, some for as much as 100 trillion years (Adams, 2008, p. 39). If multiple locations were colonized, the risk of total destruction would dramatically decrease, since it would take independent global disasters, or a cosmological catastrophe, to destroy civilization. Because of this, it is possible that our descendants would survive until the very end, and that there could be extraordinarily large numbers of them.

This scenario seems speculative and discussion of it may seem more fitting for science fiction than for serious academic philosophy. I readily acknowledge that we cannot be confident in any concrete predictions about the long-term future. At the same time, it would seem unreasonable to be highly confident that our descendants will not colonize space. After all, it is plausible that colonizing space is technologically possible, and that given one billion years of technological development, our descendants would be able to do very many of the technologically possible things they'd be interested in doing. And as we've said, if our descendants *do* colonize space, the risk of extinction becomes much lower. In light of these considerations, it is a live possibility that if our descendants do survive for the next billion years, they will colonize many stars and survive for the full 100 trillion years ahead of us. As above, this doesn't tell us anything about the *objective* probability of our descendants eventually colonizing space, but it tells us that the *subjective probability*, in the sense of "reasonable betting odds," is not extremely low. Therefore, we should assign colonization and

long term survival a subjective probability greater than 1/100, conditional on surviving for a billion years. We should therefore agree that the unconditional probability of this event is at least 1 in 10,000 (since $\frac{1}{100} \times \frac{1}{100} = \frac{1}{10{,}000}$). Therefore, there are at least $\frac{1}{10{,}000} \times 10$ trillion years = 10 billion expected years of civilization ahead of us.

## 3.2 A framework for estimating the value of a chance of a long future

The goal of this section is to outline a method of estimating the value of the future which would give approximately correct answers in some class of cases that lets us say something helpful about how good the future could be, supposing we were allowed to know everything about what happened in the future. We'll use this method, together with the conclusions of the previous section, to argue that a long future is extremely important.

I find it helpful to imagine we're designing a computer program that makes this estimate for us. What we want is a computer program that gets fed a possible history of the world, and then gives us an estimate of how good that history of the world is. We don't know exactly what this computer program should look like, but my idea is to say a few things about what it might look like, and then reach some conclusions on the basis of those assumptions.

First, the program would divide the history of the world into *periods*, chunks of time of some large duration (such as 100, 1000, or 10,000 years). Second, for each of these periods, it would look at what happens during that period. On the basis of what happens during that period, it would assign that period a score which says how good that period was. This score could be a number, but it in giving the value of a period a number, we need not assume that periods could, in principle, be given precise values. Like Parfit (1984, p. 431), we might hold that in principle, only rough evaluation is possible. Third, it would use these scores to come up with an estimate of the value of the whole history.

Obviously, you might get different answers depending on how you carve up the world into periods of time, and there is no "privileged" way of carving up history into periods. This does not concern me because I'm designing a method for making a rough approximation in a particular class of cases, rather than trying to provide a precise "final theory" of population ethics. It is not worrying if some approximation technique calls for judgment or might have been done any various ways that would give moderately different answers. Many valuable approximation techniques have these properties.

Equally obviously, it would be totally impractical to build something that could execute this computer program. Furthermore, if we were going to actually make this computer program, rather than just talk about its properties, we'd need to do a lot more to spell out each of these steps. I'm not going to try to spell out the second step much at all, and we'll see that this is a strength of my strategy. But I would like to propose some reasonable conditions for how the third step should go. These conditions imply that the third step should proceed by something like adding up the scores across periods.

### 3.2.1 Period Independence

The first, and most important, normative assumption that I will use is:

> *Period Independence*: By and large, how well history goes as a whole is a function of how well things go during each period of history; when things go better during a period, that makes the history as a whole go better; when things go worse during a period, that makes history as a whole go worse; and the extent to which it makes history as a whole go better or worse is independent of what happens in other such periods.

As I said in the first chapter, by "independent" I mean *metaphysically* independent, rather than causally independent. Obviously, what happens now can have profound causal impacts on the future. But when we ignore such causal effects, Period Independence claims that the additional value of things going better during a certain period can't depend on what happens in other periods.

It is easier to understand Period Independence if we're clearer about what it rules out. To do that, let's first consider an analogy. Some philosophers argue that how well a person's life goes for him depends on the "shape" of his life, and not just the total amount of good that he enjoys at each moment of his life. For instance, two lives might contain the same amount of good moment-by-moment, but in one of these lives the moment-by-moment well-being may increase over time, whereas in the other it would decrease over time. Some philosophers have argued that, in such cases, though each life contains the same amount of moment-by-moment welfare, the life with increasing moment-by-moment well-being is better because of its "shape" (Velleman, 2000, pp.58-59). Period Independence says that, in general, there is no such dependence on "shape" across different periods of history.

For an illustration of Period Independence, consider the following possible histories of the world:

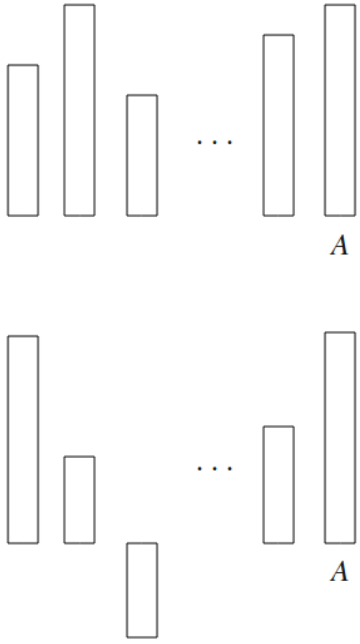


Figure 3.1: Illustration of Period Independence

In this graph, the rectangles represent different periods of history. Their width indicates their duration, and their height indicates how well things go during the period (per unit time). According to Period Independence, adding or removing period $A$ from the world's history would be equally good in either case. How good it was that period $A$ happened could not depend on how well things go during other periods.

Therefore, we can calculate the value of the whole of history by starting from the beginning and asking how much the first period contributes, and then asking how much the second period contributes, and so forth. The result is that the value of the whole of history is, approximately, the sum of the value of what happens in each period. In saying this, I don't mean to take a stand on questions like whether many very small benefits can add up to a very large benefit. My general views don't depend on that question.

### A rationale for Period Independence

To appreciate the rationale for Period Independence consider the following scenario:

*Asteroid Analysis*: World leaders hire experts to do a cost-benefit analysis and determine whether it is worth it to fund an Asteroid Deflection System. Thinking mostly of the interests of future generations, the leaders decide that it would be well worth it.

And then consider the following ending:

*Our Surprising History*: After the analysis has been done, some scientists discover that life was planted on Earth by other people who now live in an inaccessible region of spacetime. In the past, there were a lot of them, and they had really great lives. Upon learning this, world leaders decide that since there has already been a lot of value in the universe, it is much less important that they build this device than they previously thought.

On some views in population ethics, the world leaders might be right. For example, if we believe that additional lives have diminishing marginal value, the total value of the future could depend significantly on how many lives there have been in the past. Intuitively, it would seem unreasonable to claim that how good it would be to build the Asteroid Deflection System depends on this information about our distant past. Parfit and Broome appeal to analogous arguments when attacking diminishing marginal value and average views in population ethics. See Parfit (1984, p. 420) and Broome (2004, p. 197) for examples.

**Objection: Period Independence ignores some important "shape" considerations**

Some people may object to Period Independence on the grounds that how well history goes depends on averages across periods of time, how well things go at the best times, how badly things go at the worst times, whether things are getting better or worse, variety across periods of time, equality across periods of time, or shared traditions across periods of time. Defenders of Period Independence can count these holistic considerations *within* periods, but not *across* periods. I have a few responses to this.

First, we can reply that in variants of Our Surprisingly Relevant History, holistic concern for these ideals will have implausible implications, and insist that we are more certain that it would be irrational to change one's decision in Our Surprisingly Relevant History than we are that these concerns can rationally be applied across periods. By taking this strategy, we may fail to accommodate everything that certain people believe, but we might still have a more plausible view, all things considered.

Second, we can reply that decreasing existential risk and shaping the far future generally are good with respect to at least some of these concerns. For example, if we avoid premature extinction, then there is reason to believe that the human condition will improve. And thus, there is reason to believe that the peaks of human achievement lie ahead of us. Therefore, recognizing a non-period-independent value which depends on how well things go during the best times or whether things are getting better would make it more important to improve the far future. Likewise, the worst parts of human history may be in the past anyway, so a focus on the worst periods of history would neither tell significantly in favor nor significantly against shaping the far future for the better. In addition, causing human civilization to last longer seems like it would probably make things better in terms of preservation of traditions.

Third, we can claim that the future is neither expectably good nor expectably bad with respect to some of these ideals, so we don't need to worry about those ideals. In the cases of equality and variety, it is hard to tell what the relevant values are, and even harder to predict the empirical facts that are relevant to the value assessments. So it's unclear that taking account of these factors would substantially affect the value of shaping the far future.

Finally, if we decide that some of these considerations are relevant and significant for some comparisons we would like to make, we can note that as a limitation of our analysis and try to adjust for it.

### 3.2.2 Additionality

If this argument is going to work, we need to establish that:

> *Additionality*: If "standard good things" happen during a period of history—there are people, the people have good lives, society is organized appropriately, etc.—that makes that period go better than a period where nothing of value happens.

If this were not true, then a future without prosperous future people might be no worse than a future with no sentient life.

What is the intuition behind Additionality? It goes back to our "computer program" analogy above. When we write the computer program that estimates the value of a possible future, it makes sense for that computer program to look at how well things go during each period in that possible future. And it seems that in order to know how well things go during a given period, the computer program should just have to look at qualitative facts about what happens during that period, such as what kind of "standard good things" are happening during that period. All the standard good

things that are happening now are good, and better than a "blank" period where no standard good things happen. So if similar things happen in future periods, that should be good as well.

In my argument, Additionality rules out strict Person-Affecting Views. There might be other reasons people would deny Additionality, but I have no such reasons in mind. According to strict Person-Affecting Views, the fact that a person's life would go well if he lived could not, in itself, imply that it would be in some way good to create him. Why not? Since the person was never created, there is no person who could have benefited from being created. On this type of view, the only be important to ensure that there are future generations if it would somehow benefit people alive today, or people who have lived in the past (perhaps by adding meaning to their lives). If one does not accept a view of this kind, I see no reason to think that it doesn't matter whether "standard good things" happen in the future.

It would be very strange to respond to my arguments by appealing to a strict Person-Affecting View because, as I'll argue in the next chapter, these views have obviously implausible implications when considering cases involving human extinction, and cases of human extinction are central to the issues under discussion.

We can contrast strict Person-Affecting Views with *moderate* Person-Affecting Views. On these moderate views, creating people who would have good lives if they lived is good, it just often isn't as good as improving the lives of existing people by equivalent amounts. If this type of view is correct, it may be that additional future periods where "standard good things" happen might be less valuable than the current period with its "standard good things." A fairly natural way to spell this out would be to say that these future periods have only some fraction of the value that they would have had if the value were calculated in a non-Person-Affecting way. If this fraction is not unreasonably small, it will not substantially affect our conclusions about the value of shaping the far future. I'm intending to give arguments that the far future is much, much more important than what happens in the short run, and multiplying by a fraction that isn't unreasonably small is unlikely to change the big picture. In any case, I discuss these views at much greater length in the next chapter.

### 3.2.3 Temporal Impartiality

The next important assumption is:

*Temporal Impartiality*: the value of a particular period is independent of when it occurs.

This assumption is not very controversial among philosophers, but many economists reject it. On their view, we should count benefits that come in the future as intrinsically less important than

benefits that come sooner, and the value of future benefits should decrease exponentially with time. Since Parfit (1984, Appendix F), Cowen and Parfit (1992), and Broome (1992) have convincingly argued against this position and few philosophers believe it anyway, I will only briefly explain why it should be rejected.

Some rather obvious examples suggest that there is no fundamental significance to when benefits and harms take place. To take an example from Parfit (1984), suppose I bury some broken glass in a forest. In one case, a child steps on the broken glass 10 years from now and is injured. In another case, a child steps on the broken glass 110 years from now and is injured in precisely the same way. If we discount for time, then we will count the first alternative as much worse than the second. If we use a 5% discount rate per year, we should count this alternative as over one hundred times worse. This is very implausible.

Economists often appeal to two arguments for pure temporal discounting. First, they argue that people want to discount the future and governments should echo the will of the people. As Parfit (1984, Appendix F) points out, there is a difference between the questions:

1. If people decide to do $A$, what should governments do?

2. Should people decide to do $A$?

I am arguing about what people ought to decide, so the democratic argument has no clear relevance.

Second, some economists argue that if we do not discount future benefits, we would have to spend far too much of our resources on future generations (Posner, 2004). This assumes that we are required to do whatever is best. It is a familiar fact that if we are required to do whatever is best, life will be demanding for us. That's a classic objection to the view that we are required to do whatever is best. If we want to avoid these demands, we should revise the view that we are required to do whatever is best, perhaps placing some limits on how much we should be required to sacrifice to promote good outcomes. We should not make implausible claims about the comparative badness of exactly when children step on glass in forests.

### 3.2.4 Risk Neutrality

My argument relies essentially on expected value considerations. I assume:

> *Risk Neutrality*: The value of an uncertain prospect equals its expected value.

This assumption is important because, in all probability, any given project will do very little to affect the long-term prospects of civilization. Therefore, my argument must proceed by arguing that

the value of the future is extremely large, so that reducing existential risk by a small probability, or having some small probability of creating some other positive trajectory change, is also very large. The most straightforward way to do this is to use the Risk Neutrality assumption to argue that reducing existential risk by some fraction is as important as achieving that fraction of the potential value of the future.

### 3.2.5 Objection: don't these assumptions entail the Repugnant Conclusion?

They don't, but they eliminate one kind of strategy for avoiding it.

According to the Repugnant Conclusion:

> For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better, even though its members have lives that are barely worth living. (Parfit, 1984, p. 388)

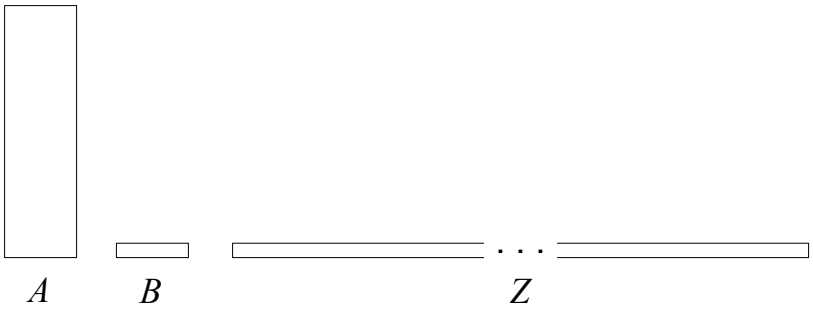


Figure 3.2: Period Independence and the Repugnant Conclusion

Here is why one might be misled into believing that Period Independence entails the Repugnant Conclusion. By Additionality, $B$ has some positive value, call it $y$. If we string together a long enough sequence of periods like $B$, we will get a world like world $Z$. Since each period contributes an equal amount of value, the value of $Z$ must be some very large multiple of $y$. Therefore, one might think, Period Independence entails that $Z$ is better than $A$, provided $Z$ just has enough periods.

The crucial issues with this argument are whether $B$ is good at all and whether adding up a lot of small benefits can lead to something arbitrarily good. I haven't taken a stand on these

questions. $B$ might be bad, even if Additionality, Period Independence, and Temporal Impartiality are true. For a proof of this, consider critical-level utilitarianism. According to the sort of critical-level utilitarianism defended by Broome (2004) and Blackorby et al. (2005), it can be a bad thing for some people to exist, even if their lives are worth living. On their view, rather than adding up total well-being to determine how good an alternative is, we should add up the extent to which each person's well-being exceeds a certain threshold. This view is compatible with Additionality, Period Independence, and Temporal Impartiality, but it avoids the Repugnant Conclusion.

Even if $B$ is good, my assumptions might be true while the Repugnant Conclusion is false. We might deny that many periods that have very little value can "add up" to something that is arbitrarily good. A ranking of this kind may be represented lexicographically and be made consistent with Period Independence, avoid the Repugnant Conclusion, and still let many large benefits add up to an amazingly large benefit (which is what I need for my arguments).

Therefore, my assumptions do not *commit* us to the Repugnant Conclusion. True, lexicographic and critical-level views have many implausible implications. But that's par for the course in population ethics. As Parfit (1984), Arrhenius (2013), Temkin (2012), and others have shown, there will be great problems however we try to avoid the Repugnant Conclusion, so we should not be too quick to dismiss these approaches.

Moreover, denying Additionality and accepting some strict Person-Affecting View is not an effective way to deal with the Repugnant Conclusion in any case. Suppose that we must decide whether the future will look like $A$ or like $Z$, and that none of the people we are considering exist yet or will exist regardless of what we do. Rather than implying that $A$ is better than $Z$, as desired, a strict Person-Affecting View implies that there is nothing to choose between $A$ and $Z$.

What's true is that my argument rules out one class of ways of avoiding the Repugnant Conclusion: theories with diminishing marginal value of population across periods (including average views). These three strategies (lexical views, critical-level theories, and theories of diminishing marginal value) cover most of the remotely plausible ways of avoiding the Repugnant Conclusion, within a broadly welfarist axiology. Since I am only taking one strategy off the table, it is not true that my assumptions entail or inevitably lead to the Repugnant Conclusion.

## 3.3 What do these assumptions suggest about the value of shaping the far future?

### 3.3.1 How valuable is the far future, assuming it goes well?

Our assumptions imply that we can approximate the value of the far future by dividing the future of the world into periods, assigning value to each period in a way that is independent of when it happens, and adding up the value across the periods. To see where this leads, consider a possible history of the world where humanity survives for 100 trillion years as outlined in section 3.1, and suppose we divide that history up into a trillion 100 year periods. Then the whole history is a trillion times as valuable as what happens during an average 100-year period. If we divided periods into a million years, the whole history would be 100 million times as valuable as one average million-year period. I prefer to think in terms of the 100-year period, since I feel I have a better grip on what that's worth. We could run similar arguments with other proposed durations for periods.

In a world where humanity survives for an enormously long time, should we expect future periods to be, on average, better, worse, or about equally as good as current periods? I'm inclined to think that if our descendants managed to create a vast civilization and they preserved a decent portion of the good aspects of our values, future periods would be, on average, much better than current periods. But even if average future periods were only about equally as good as the current period, the whole of the future would be about a trillion times more important, in itself, than everything that has happened in the last 100 years.

### 3.3.2 How valuable is the far future, in light of our uncertainty about how long it will last?

To determine the value of a chance of humanity surviving for a very, very long time, we have to multiply by the chance of that happening. That tells us what the opportunity cost of premature extinction is, and it helps us determine how important it is change our trajectory for the better. In section 3.1, I argued that it is reasonable to assign at least a 1% probability to the possibility that humanity survives for 1 billion years, and at least a 1% probability to the possibility that humanity survives for 100 trillion years, given that humanity survives for 1 billion years. Therefore, the expected duration of humanity's existence is at least $1\% \times 1\% \times 100$ trillion years = 10 billion years. And again, assuming that future periods are expected to go at least as well as the current 100-year period, that's at least 100 million times more important, in itself, than everything that

happened in the last 100 years.

### 3.3.3 How valuable is existential risk reduction in comparison with proximate benefits?

In this section, I'll argue that, on the level of global priorities, existential risk reduction is much more important than producing proximate benefits, in the sense that the opportunities to do good are much, much greater. A major qualification of this claim is that I don't mean to argue that benefits from feasible ways of reducing existential risk are much better than feasible ways of producing proximate benefits *once the ripple effects of proximate benefits are included*. I'll return to this point below.

On a global level, what could feasibly be done to provide proximate benefits, and what could feasibly be done to reduce existential risk? A dramatic victory around the world might make this period go *twice* as well as it otherwise would, say. What kind of existential risk reduction would be required to produce comparable benefits? Given our assumptions from the above section, decreasing the probability of a particular risk by one in a million would result in an additional 10,000 expected years of civilization, and that would be at least 100 times better than making things go twice as well during this period (not counting ripple effects, a qualification I leave implicit hereafter).

It is not hard to believe that, collectively, humanity could do things that would decrease the risk of some existential catastrophes by one in a million. One major reason to believe this is that we've recently done a number of things that have reduced existential risk. We've made it through the cold war and scaled back our reserves of nuclear weapons. We've tracked most of the large asteroids near Earth, so that we'd probably be able to respond if one were on track to collide with Earth. We've built underground bunkers for "continuity of government" purposes, which might help humanity survive certain catastrophes. We've instituted disease surveillance programs which would allow the world to respond more quickly in the event of a large-scale pandemic. We've identified climate change as a potential risk and developed some plans for responding, even if we've done rather little so far. We've also built institutions that reduce the risk of extinction in subtler ways, such as decreasing the risk of war or improving the government's ability to respond to a catastrophe. For more detail on the general decline of violence, see Pinker (2011).

Another reason to believe that we could reduce existential risk by one in a million is that many of these efforts could be improved. We could track more asteroids, build better bunkers, improve our disease surveillance programs, reduce our greenhouse gas emissions, encourage non-proliferation

of nuclear weapons, and strengthen world institutions in ways that may further decrease existential risk. There is still a substantial challenge in identifying worthy projects, but it seems likely that such projects exist.

To sum up, relatively small reductions in existential risk are much more important, in themselves, than very large proximate benefits. This makes it plausible that reducing existential risk is, in itself, a more important goal than providing proximate benefits, though this comparison does not include potentially important ripple effects of providing proximate benefits.

### 3.3.4 How valuable is existential risk reduction in comparison with speeding up development?

For similar reasons, it is plausible that existential risk reduction is more important, in itself, than speeding up development. As we saw in the above section, it is not unrealistic to consider scenarios in which humanity reduces existential risk by one in a million, and this results in an additional 10,000 expected years of civilization. In contrast, the amount that we could realistically speed up humanity's technological and moral progress in this period is much more modest, probably measured in decades at best. Because of this, it's plausible that existential risk reduction is more important than speeding up development, provided we ignore ripple effects from speeding up development, a qualification I'll return to below.

### 3.3.5 Why "focus on trajectory changes," rather than "minimize existential risk" is the upshot of this discussion

I've now argued that proximate benefits and benefits from speeding up development are less important than benefits from reducing existential risk. Someone might argue, on this basis, that existential risk reduction is the most important way of shaping the far future. And, as I mentioned in the introduction, Bostrom (2012) has made roughly this argument. He concluded:

> "[T]he loss in expected value resulting from an existential catastrophe is so enormous that the objective of reducing existential risks should be a dominant consideration whenever we act out of an impersonal concern for humankind as a whole. It may be useful to adopt the following rule of thumb for such impersonal moral action:
>
> *Maxipok:* Maximize the probability of an "OK outcome," where an OK outcome is any outcome that avoids existential catastrophe." (Bostrom, 2012, p. 10).

This conclusion, however, does not follow because there may be other ways to have a large, persistent effect on the far future without reducing existential risk. Bostrom recognizes that it is possible for the future to be significantly flawed without human extinction, so it's worth emphasizing that he defines an existential catastrophe to include not only humanity's extinction, but also "the permanent and drastic destruction of its potential for desirable future development," (Bostrom, 2012). But as I said in section 1.1.2.3, there could be many positive or negative trajectory changes which would not be *drastic* curtailments of humanity's future potential. Some persistent changes in values and social norms could make the future one hundredth, one thousandth, or one millionth better or worse, without there being any drastic changes to the far future. And I see no clear reason why we should expect existential risks to be more worthy of our focus than these other trajectory changes. Sure, succeeding in preventing an existential catastrophe would be better than making a smaller trajectory change, but creating a small positive trajectory change may be significantly easier.

I do think my arguments support a more general rule of thumb: what matters most for shaping the far future is producing positive trajectory changes and avoiding negative ones. This is more general because preventing an existential catastrophe is one kind of trajectory change. It's supported by my arguments because

1. The categories of "proximate benefits," "benefits from speeding up development," and "benefits from trajectory changes," appear to cover the most important categories for shaping the far future, and

2. I've already argued that one class of trajectory changes, existential risk reduction, is more important than providing benefits from speeding up development and proximate benefits.

### 3.3.6 How important are ripple effects?

Ripple effects from ordinary actions can be very significant. It would be a mistake to read the above arguments and conclude that actions targeted at creating positive trajectory changes are essentially always better than actions which only have ripple effects on the far future because:

1. Many actions which provide proximate benefits speed up development.

2. Speeding up development may have non-negligible indirect effects on existential risk or other trajectory changes.

This may not be the most important type of ripple effect. But I found it to be the easiest one to think about, so I'll discuss it in this section.

In defense of the first premise, lots of ways of benefiting people make them happier and more productive. True, some people are not productive members of society, and helping them may have limited ripple effects, but this is not the general case. But it is often hard to predict in advance who these people will be, so helping people in general has some expected impact on people's productivity, and that has indirect effects on things like the rate of cultural and technological progress.

How could speeding up development affect existential risk? Some "natural" existential risks may be reduced by speeding up development. For example, the risk of a 10-km asteroid hitting the earth in a given century has been estimated at approximately one in a million (NASA, 2007; Chapman, 2004). Many scientists believe that an asteroid impact of this size caused the extinction of the dinosaurs. An asteroid of that size hitting the earth could potentially result in human extinction. Tracking near-earth asteroids has substantially reduced this risk, since we now know the positions of the vast majority of large near-earth asteroids, and if an asteroid were known to be on a collision course with earth, it would probably be possible to deflect it. Past events that sped up technological progress in general made these asteroids get tracked sooner, and thereby reduced existential risk.

For another example, speeding up development may reduce the probability of a critical resource shortage. It may be that certain resources are essential for reaching a higher level of technological development, but less necessary later. If we make technological progress faster, we may be less likely to run out of those resources before reaching the critical development threshold. Even if reaching this level of development is not essential for preventing human extinction, failing to reach it could carry an enormous opportunity cost, such as failing to colonize other planets.

Speeding up development may also reduce some anthropogenic existential risks, such as a nuclear war. If having a long future is possible at all, then it must be possible for humanity to reach a state where the risk of total destruction in a given century is extremely small. This might be because very effective surveillance technology develops, or because very effective political institutions are built. In this type of scenario, speeding up development would make that technology or those institutions arrive sooner, reducing existential risk.

Because of this, the effects of speeding up development, in itself, may be swamped by its effects on existential risk. Here's a speculative back-of-the-envelope calculation that illustrates the point. Suppose that in the next 100 years, there is a 1% chance of a war between major nuclear powers, a 1% chance that this leads to a major nuclear war, and that there is a 1% probability of human extinction given a major nuclear war. Suppose that if we moved through this period in 99 years rather than 100, the risk of a nuclear war between major powers would be 1% less. Then speeding up development by 1 year would reduce existential risk by $1\% \times 1\% \times 1\% \times 1\% = 10^{-8}$, which would

result in an additional $10^{-8} \times 10^{10} = 100$ expected years of civilization.

There could be other reasons that speeding up development would increase existential risk. Perhaps someone else could make their own back-of-the-envelope calculation which suggests that speeding up development by 1 year would result in 100 fewer expected years of civilization. I'm open to that. What seems unlikely is that all of this "cancels out," so that effects of speeding up development on existential risk is negligible. Instead, it seems more likely that the expected effects of speeding up development on existential risk, whether positive or negative, are substantial.

The upshot of all this is that even if my arguments are accepted, ordinary actions may be very important because of their effects on the far future. Thus, even if one accepted a very extreme version of my thesis that shaping the far future is overwhelmingly important, we could not ignore or discount the effects of ordinary actions on the far future. This is a helpful reminder that, as I said in chapter 1, the best ways of shaping the far future may be very broad.

This upshot does not take the teeth out of my claim that shaping the far future is overwhelmingly important. Currently, we have that what matters most is creating positive trajectory changes. We don't know whether the best way to do this is very broad or very targeted. If it is very targeted, then my conclusion may have very revisionary implications for our understanding of what matters most because few of us think about good accomplished primarily in terms of trajectory changes. If it is very broad, the implications would be less revisionary, but still potentially significant. To use an example from chapter 1, ripple from saving lives in poor countries may be substantially smaller than the ripple effects of saving lives in rich countries, though this consideration wouldn't normally get much weight in deciding which lives to save. In general, if we accepted this line of reasoning, it would become more important to think much more carefully about ripple effects. And some moderately broad, moderately targeted ways of shaping the far future, such as promulgating norms that emphasize the importance of future generations, may be much more important than common sense would say they are. In any case, since we do not currently know how targeted the best ways of creating positive trajectory changes are, and we don't currently have a good understanding of how to create the most significant ripple effects, at the very least, my arguments suggest that certain lines of inquiry related to these questions are much more important than we would have thought in advance.

## 3.4 Conclusion

I have presented some normative and empirical considerations in favor of the conclusion that shaping the far future is overwhelmingly important. The key claims are that humanity could survive for a very long time, with an expected duration on the order of billions of years or more; that the future is overwhelmingly important if my normative assumptions are true; that we could potentially shape the future for the better by speeding up progress, reducing existential risk, or producing other positive trajectory changes; and that what matters most for shaping the far future is creating positive trajectory changes. The best ways of shaping the far future could be very broad or very targeted, and knowing which would be very valuable.

There are many remaining questions about both my normative assumptions and my speculation about the far future. The coming chapters will focus on the normative questions in much more detail.

# Chapter 4

# Should "Extra" People Count for Less?

## Introduction

In the previous chapter, I argued that shaping the far future is overwhelmingly important. There are two major positions which, if true, would undermine this view. Both views can be understood as departures from simple positions like Total Utilitarianism. Total Utilitarians determine the value of an alternative by adding up the total well-being of all the people that would ever exist if that alternative were chosen. The first kind of revision involves tinkering with the *form* of the aggregation procedure (changing from summing over people to, for instance, giving additional people diminishing marginal value). The second kind of revision involves tinkering with the *scope* of the aggregation procedure (changing from aggregating over all people that would ever exist if the alternative were chosen to aggregating over some subset of those people). The idea behind restricting the scope of the aggregation is to allow us to ignore (or discount) any supposed non-instrumental value in creating "extra" people. I therefore call the first kind of view a *theory of diminishing marginal value*, and the second kind of view is a *Person-Affecting View*. In this chapter, I will only be concerned with Person-Affecting Views.[1]

A Person-Affecting View can be characterized in terms of which people it counts as "extra." I use scare quotes around "extra" because the term is really just a place-holder for a certain kind

[1] When I say that these views can be understood as departures from Total Utilitarianism, I do not mean to implicate that they must be understood in this way. We could have a Person-Affecting version of egalitarianism, prioritarianism, perfection, maximin, or any combination of these views. One could also combine *both* revisions, though I will not discuss this possibility.

of restriction; I have no interest in entering debates about the semantics of "extra." On most of these views, different people are counted as "extra" at different times or choice situations. Which people count as "extra" for the purposes of comparing two alternatives may also depend on which alternatives are being compared. Some prominent examples of these views are as follows:[2]

| Name of View | Which people are "extra"? |
|---|---|
| Presentism | The people that do not presently exist |
| Actualism | The people that will never actually exist |
| Necessitarianism | The people whose existence is dependent on which (of perhaps many) alternative is chosen |
| Comparativism | The people that exist in one alternative being compared, but not the other |

Table 4.1: Classification of Person-Affecting Views

The Person-Affecting View claims:

> *Person-Affecting View*: When aggregating the interests of different people to determine the value of an outcome, the interests of "extra" people count for less or can be ignored.

On *asymmetric* versions of these views, people who would have lives that would not be worth living are never considered "extra."

A Person-Affecting View may be either *moderate* or *strict*. On a strict version, the interests of the extra people are not weighed in the aggregation procedure. On a moderate version, their interests are weighed, but they are given less weight. This might be accomplished, for instance, by changing Total Utilitarianism so that these people's interests are discounted by a constant factor. Other variations would be possible as well.

Many authors distinguish between a wide and narrow version of the Person-Affecting View, and on the wide reading, the interests of "extra" people are not discounted. People with Wide Person-Affecting Views disagree with people with so-called "impersonal" views because they disagree about *why* the interests of "extra" people matter. I will not be concerned with the distinction between Wide Person-Affecting Views and impersonal views; these are both examples of what I call *Unrestricted Views*, since they both deny that "extra" people count for less. The main claim I want to defend in

[2]This terminology is borrowed from Arrhenius (2000, chapter 8).

this chapter is that Unrestricted Views are more plausible than any of the Person-Affecting Views that I consider. Toward the end of the chapter, I also argue that it is unlikely that any Person-Affecting View developed in the future will undermine the argument that shaping the far future is overwhelmingly important.

I will begin by considering the arguments for Person-Affecting Views. One may try to argue for a Person-Affecting View by appeal to theoretical considerations or intuitive considerations. In the first section, I will consider two theoretical arguments for Person-Affecting Views which, I argue, are not compelling. In the second section, I consider whether intuitive considerations favor accepting the Person-Affecting View. I consider Asymmetric Person-Affecting Views first. I argue that while these views can capture some seemingly important intuitions about some cases, they have very implausible consequences elsewhere, especially in cases of extinction. Next, I consider Moderate Person-Affecting Views. I argue that these views do better than strict views, and avoid many of the extinction problems. However, these views stand in an uncomfortable position. If the weight given to "extra" people is too high, they behave too similarly to Unrestricted Views. If the weight is too low, they face many of the challenges of strict views. In the concluding section, I take stock of the objections given thus far and introduce a related principle, which I call the Personal Good Restriction. The Personal Good Restriction is an Unrestricted View that fits well with some important theoretical motivations behind the Person-Affecting Argument, but it avoids all of the difficulties of Person-Affecting Views. I argue that, all things considered, the objections to this view are not nearly as powerful as the objections to Person-Affecting Views. I conclude that this view is more plausible than any of the Person-Affecting Views I have considered.

**Limitations of my analysis**

This will not be a comprehensive review and critique of all Person-Affecting Views. That would require a dissertation in itself. There are some important views I do not think fall under any of the categories I have discussed, including views which claim it is indeterminate whether creating additional happy people is neutral or good; an interesting variation on which people count as “extra” due to Meacham (2012); some interesting unpublished work by Ralf Bader; and probably other work I don’t know about. There are also some important examples, such as the slave child case due to Kavka (1982), that I have not discussed. I think the examples I do discuss cover most of the important test cases in this literature, and I think that most of my arguments carry over, in some form or other, to the Person-Affecting Views that I do not discuss.

## 4.1 Theoretical arguments for Person-Affecting Views

### 4.1.1 The Person-Affecting Argument

#### Argument presentation

Let's have another look at the Person-Affecting Argument:

> *Person-Affecting Restriction*: If one alternative is better than another, then that alternative is better for someone. If one alternative is better for someone than another alternative and it is worse for no one, then that alternative is better than the other.
>
> *Incomparability of Non-existence*: If a person exists in one alternative but not the other, no positive comparative claim about his well-being in these alternatives (such as the claim that he is better off, equally well off, worse off, or roughly equally well off in one rather than the other) is true.
>
> Therefore, an outcome in which an additional happy person exists, but no one else is better or worse off, is not better than an outcome in which this person does not exist.

To illustrate this argument, consider a version of an example from Parfit (1984, p. 381):

> *The Happy Child.* Some parents are deciding whether or not to have an additional child. On balance, having the additional child would be neutral for others, and neutral for them. But the child would have a happy and meaningful life if he lived.

By Incomparability of Non-existence, the child could not be better off than he would be if he were never born. So having the child is better for no one. Therefore, by the Person-Affecting Restriction, there is no way in which it would be better if the parents had this child.

#### Are the premises compelling?

One finds the earliest version of the Person-Affecting Argument in Narveson (1967). The first premise stems from the idea that any sense in which one outcome is "better" than another must be reducible to the outcome being better for particular people. If someone claims that one alternative is worse than another, though choosing it would mean no loss for anyone, his claim is, *prima facie*, a puzzling one. "Worse for whom?" we might rhetorically ask.

As Temkin (2000) points out, we can see much of the plausibility of the Person-Affecting Restriction by seeing how readily it is invoked in important arguments in philosophy and economics. The Person-Affecting Restriction can explain why there is no value in implementing a government

policy that benefits no one, and no value in stopping people from doing things that harm no one. It explains why many people find it compelling to think that Mere Addition cannot make an alternative worse: it involves making no one worse off. (By "Mere Addition", I mean transforming population $A$ into population $A+$ in the diagram below).[3]

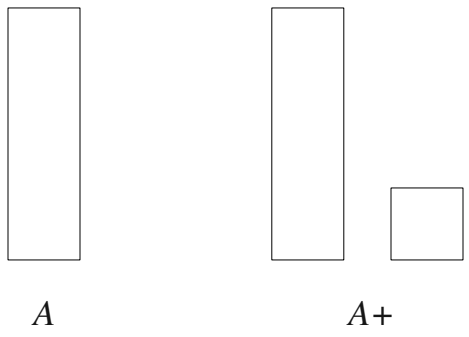


Figure 4.1: Mere Addition

It can also explain why, as many believe, there is nothing good about Leveling Down. Although many people seem to value equality of well-being in some contexts, the Leveling Down argument suggests that equality is not intrinsically valuable. Consider the move from $C$ to $D$ in the diagram below.

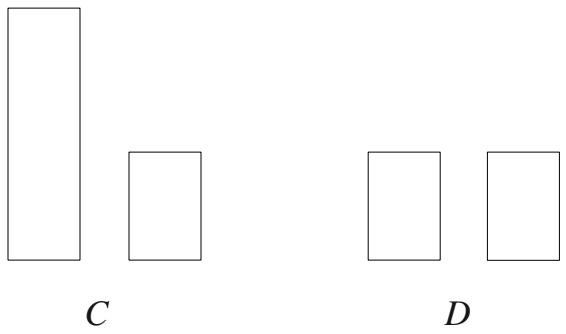


Figure 4.2: Leveling Down

[3]In this chapter, the width of a column indicates the number of people living in a group, and the height indicates the well-being of the people in that group. These diagrams are not snapshots, but representations of well-being distributions of over the whole history of a possible world.

Because Leveling Down is worse for someone and better for no one, many find it intuitive that $D$ is in no way better than $C$, even though $D$ has greater equality (Parfit, 1997).[4]

The justification for the second premise, Incomparability of Non-Existence, is usually offered in metaphysical terms. Broome expresses the thought clearly:

> If it were better for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been worse for her than if she had lived. But if she had never lived at all, there would have been no her for it to be worse for, so it could not have been worse for her. (Broome, 1999, 168)

**Problems with the Person-Affecting Argument**

Having explained the argument, let's examine its problems. The most troubling issue is that the argument delivers a standard Person-Affecting View, rather than an asymmetric one. To see this, let's consider one of Parfit's cases:

> *The Wretched Child:* Some woman knows that, if she has a child, he will be so multiply diseased that his life will be worse than nothing. He will never develop, will live for only a few years, and will suffer from pain that cannot be wholly relieved. If she has this child, it will not be good or bad for anyone else.

Of this case, Parfit says, "Even if we reject the phrase "worse than nothing," it is clear that it would be wrong knowingly to conceive such a child." (Parfit, 1984, p. 391). Parfit is surely right about this. However, Incomparability of Non-Existence entails that non-existence cannot be better for the child than living this wretched life. By stipulation, the child's existence affects no one else, so the child's existence cannot make the outcome worse. Thus, the Person-Affecting Restriction implies that it cannot be worse if the Wretched Child exists.

So it seems that the Person-Affecting Argument proves too much; what we want is an Asymmetric Person-Affecting View, and this argument isn't going to give it to us. Might we change the argument, perhaps in a subtle way, so that it yields an Asymmetric Person-Affecting View?

One might try to claim that while existence can be worse than non-existence, it cannot be better. Though one can't rule it out in advance, this position seems arbitrary. To be sure, it does not fit

[4]Temkin (2000) discusses these and other applications of the Person-Affecting Restriction in the context of motivating the position before he attacks it. As we will see below, some views that may violate the Person-Affecting Restriction, such as Total Utilitarianism, accommodate these judgments as well.

with the metaphysical flavor of the argument for Incomparability of Non-Existence. We can see this by replacing every occurrence of "better" with "worse" (and vice versa) in the quote from Broome:

> If it were [worse] for a person that she lives than that she should never have lived at all, then if she had never lived at all, that would have been [better] for her than if she had lived. But if she had never lived at all, there would have been no her for it to be [better] for, so it could not have been [better] for her.[5]

There seem to be no metaphysical grounds for accepting an asymmetric version of Incomparability of Non-Existence, so, at the very least, we need a quite different argument if we want to establish an Asymmetric Person-Affecting View.

### 4.1.2 The Victim Requirement[6]

Another theoretical justification for Asymmetric Views might appeal to the following asymmetry:

1. If we do not create the Happy Child, the Happy Child could not be a victim (because he wouldn't exist).

2. If we create the Wretched Child, the Wretched Child would be a victim.

When I say that some person is a victim in some outcome, I mean something very vague and general. Roughly, someone is a victim if they can make a justifiable personal complaint about the outcome, such as "I had a bad life" or "I would have been a lot better off if I'd gotten an education." Roberts (2003, p. 162) appeals to this type of difference in order to defend an Asymmetric View. It is the absence of a victim, she argues, that makes it in no way good to create the Happy Child. Parfit (1984, p. 526) suggests that a view in this family would be the best vindication of an Asymmetric View, if such a view did not suffer from other objections.

To be plausible, this justification would have to explain why the above asymmetry was relevant, whereas this other asymmetry is not:

1. If we do not create the Wretched Child, the Wretched Child could not be a beneficiary (because he wouldn't exist).

2. If we create the Happy Child, the Happy Child would be a beneficiary.

[5] Roberts (2003) makes this point explicitly.

[6] I got the idea of "The Victim Requirement" from some unpublished notes for a course on population ethics taught by Jeff McMahan at Rutgers in 2012.

When I say that some person is a beneficiary in some outcome, I mean the opposite of a victim. Roughly, someone is a beneficiary if they can make a justifiable personal anti-complaint (we surprisingly don't have a word for this) about the outcome, such as "I had a good life" or "I'm a lot better off because I got an education."

To cut to the chase, I don't see any theoretical reason why the first asymmetry would be relevant but the second would not. Intuitively, both facts about victims and facts about beneficiaries would be relevant to determining the goodness of an outcome. Someone could come along and argue that if we just focused on victims or we just focused on beneficiaries, we could develop a view that had more plausible implications about cases. But notice that on a very broad construal of complaints and anti-complaints, a focus on complaints and a focus on anti-complaints could have the same implications in all cases that involve the same people. Because of this, there's no reason to think that either view will be more or less plausible in the cases that don't involve "extra" people. To see why, consider the following example:

| | Bob's well-being | Tom's well-being |
|---|---|---|
| $A$ | 5 | 15 |
| $B$ | 20 | 10 |

If we choose option $A$, Bob can complain that he has 15 less units of well-being than he otherwise would and Tom has no complaint. If we choose option $B$, Bob has no complaint and Tom can complain that he has 5 less units of well-being than he otherwise would. Now consider the outcomes in terms of anti-complaints. If we choose $A$, Bob has no anti-complaint, and Tom can anti-complain that he has 5 more units of well-being than he otherwise would. If we choose $B$, Bob can anti-complain that he has 15 more units of well-being than he otherwise would, and Tom has no anti-complaint. It's a symmetric situation in multiple ways:

| | Total complaints | Total anti-complaints | Largest complaint | Largest anti-complaint | Smallest complaint | Smallest anti-complaint |
|---|---|---|---|---|---|---|
| $A$ | 15 | 5 | 15 | 5 | 0 | 0 |
| $B$ | 5 | 15 | 5 | 15 | 0 | 0 |

Because of these symmetries, I see no reason to expect that thinking about the outcomes in terms of complaints rather than anti-complaints would have different implications about cases that don't

involve different people.

These views do have different implications about different people cases. And one might be more or less plausible than the other. But to assess that question, it would be better to simply look at how intuitive different types of Person-Affecting Views are. And that's what we're doing in the next section.

## 4.2 Intuitive considerations

In this section, I will consider some intuitive arguments for adopting certain kinds of Person-Affecting Views, and then consider some problems that these views face. Most of the material I am presenting here is not new. I am more focused on reviewing existing arguments than creating new ones, though a couple of the cases I use are, I believe, new. The first part will involve a discussion of Asymmetric Strict Person-Affecting Views. According to these views, it is bad to create people who would have miserable lives, but not good to create people who would have good lives. In the second part, I will discuss Moderate Person-Affecting Views, both Symmetric and Asymmetric versions. According to these views, it is good for "extra" people to have good lives, but it is not as important as it is that other people have good lives.

### 4.2.1 Asymmetric Person-Affecting Views

In this subsection, I discuss some intuitive judgments that motivate adopting a Person-Affecting View, and how well Asymmetric Person-Affecting Views accommodate these judgments. These judgments have to do with the ethics of having children, as in cases like The Happy Child, the relative priority of making people happy vs. making happy people, and the Repugnant Conclusion. Most of these cases have been considered before.

#### 4.2.1.1 Favorable cases

Let's start with the cases that these views seem to get right. It is very plausible that people are typically not obligated to have children. But some philosophers have argued that if the we do not accept a Strict Person-Affecting View, then people are obligated to have children. However, there is no clear implication from "it would be good for there to be additional happy people" to "people are typically obligated to have children." Why not? At least for people who don't already want to have additional children, it would be very demanding to ask people to have additional children. Moreover, even on a view that gives a lot of weight to creating additional people, having additional

children doesn't seem like a particularly effective way of doing good in the world in comparison with things like donating money and time to charity. So it would be strange if people were obligated to make potentially significant sacrifices in order to do something that actually wasn't all that effective as a method of doing good.

A related argument appeals to the claim that, intuitively, it isn't good in itself for there to be additional happy people. My reply is that this is not a robust intuition. It is hard to ask questions that isolate this intuition specifically, and I believe many, and perhaps even most, ordinary people would believe that having children that live meaningful lives is good in itself, apart from benefits to the parents or the rest of society. Putting this together with the last paragraph, it is not clear whether Strict Person-Affecting Views have more plausible implications about ordinary cases of procreation than Unrestricted Views.

Person-Affecting Views are more plausible in cases where we must choose between helping people that already exist and producing extra people.

> *Sight or Paid Pregnancy*: A wealthy philanthropist is considering how to spend his money. His first option is to pay for surgeries for blind people in the US. With his donations, he will restore the sight of ten people. His second option is to pay certain couples to have children (who otherwise would not have done so). As a result, ten children with good lives will be born.

If he rejects Person-Affecting Views, the philanthropist might well conclude that the second option is better. He could say:

> If I help the blind, they will gain sight. If I pay the couples to have children, the children will gain sight along with all of the other benefits that a good life has to offer. Therefore, paying these couples to have children is better than restoring sight to ten people.

Intuitively though, this is implausible. This conclusion is harder (though not impossible) to avoid on Unrestricted Views.

Some philosophers seem to believe that an Asymmetric Person-Affecting View could help explain what is wrong with Repugnant Conclusion.[7] In one way, this is true. An Asymmetric Person-Affecting View would entail that we should not make great sacrifices so that very many people can have lives that are barely worth living. In a more important way, Asymmetric Person-Affecting Views seem not to help at all with the Repugnant Conclusion. We might also want to know which future

[7] See, for instance,Temkin (1987, 152).

would be better for humanity: a future in which there are about 10 billion of us living excellent lives during any given generation, or a future in which there is a much larger number of us with lives that are barely worth living. Standard Person-Affecting Views have nothing to say about this question since both alternatives involve "extra" people. As Parfit (1984, p. 395) and Arrhenius (2013) have argued, we therefore cannot appeal to Person-Affecting Views to solve the more troubling problems of population ethics. (Of course, that these views do not solve these problems is not a reason to reject them, provided we have other reasons for being interested in them).

In sum, Asymmetric Person-Affecting Views help us say some plausible things about the ethics of having children and the relative priority of creating people vs. helping the needy, but they do not significantly help us avoid the Repugnant Conclusion. Let us turn to cases that are problematic for this view.

#### 4.2.1.2 Unfavorable cases: single person cases

In other cases, Asymmetric Views have implications that are less plausible. The most obvious problem is that Strict Asymmetric Views cannot explain why, when choosing which of two “extra” people to create, it is better to create someone who would have an excellent life rather than someone who would have a pretty good life. Since the interests of all “extra” people are ignored, there is nothing to choose between these alternatives.[8] But this is very implausible.

Though many people think that Asymmetric Views can best capture our thinking about the morality of having children, this is not true. According to common sense, it is not bad to have children under ordinary conditions, provided one can be reasonably confident that one’s child will have a good life, one can fulfill one’s duties to the child, and having the child does not interfere with one’s pre-existing obligations. But if we accept a Strict Asymmetric View, if we create a happy child, we do something that is not good. However, if we create a person with a bad life, such as the Wretched Child, we do something bad.

If some action could be bad, but could not be good, then it must be bad (in expectation). This point has enough independent plausibility, but let me illustrate it with an example. Suppose that pressing a certain button could result in some very bad outcome that we have reason to avoid (such as causing some person to have a painful disability), but could not result in anything good. At best, pressing this button would leave things as they are. If this were true, then it would be bad (in expectation) to push the button. If we accept a Strict Asymmetric View, then having a child cannot be good, but could be bad (ignoring benefits to the parents or others). Someone might claim

[8] This has been pointed out by Broome, Parfit, and many others.

that considering the welfare of the child can only count against having children, but it can be good and permissible to have children for other reasons, such as the for the sake of the parents. Though this view is consistent, it is a fairly disturbing view; it does not seem appropriate for a parent to say, in effect, "Sure if you look at my child's welfare, it's a bad call, but I'm going to do it because I personally find it fulfilling."

We can make another objection along these lines:

> *The Risk-Averse Mother*: A mother faces two options for her (as of now) merely potential child. The first option involves some risk. On this option, there is a 99.99% chance that her child will have a rich and flourishing life, but a 0.01% chance that the life will not be worth living. The second option involves no risk. On this option, the child will definitely have a life that is neutral.

| | 99.99% | 0.01% |
|---|---|---|
| Option 1 | Rich and flourishing life (neutral) | Not worth living (bad) |
| Option 2 | Neutral life (neutral) | Neutral life (neutral) |

Intuitively, the first option is much better. However, on a Strict Asymmetric View, the first option has a potential downside (since the child might have a bad life) and no potential upside (since benefits to "extra" people are ignored). On the other hand, the second option will be neutral either way. Therefore, on a Strict Asymmetric View, the second option must be better.

It might be replied that it is permissible to have a child if the child's *prospects* are sufficiently good, even if there could be no upside to having a child. Rather than appealing to the child's well-being in outcomes, someone might try to apply the Asymmetric View to prospects, claiming that there is nothing good about creating people with good prospects, but there is something bad about creating people with bad prospects. Let's call this the *Tricky Expectation View*. This view does not have the implausible implication that it is wrong to have children or that the Risk-Averse Mother should choose Option 2.

However, the Tricky Expectation View is hard to accept because it is not consistent with the principle that it is bad (in expectation) to do things which could be bad and couldn't be good. After all, if the gamble turns out badly, and a child lives with great suffering, defenders of the Tricky Expectation View will count this as a bad result, though they will count the production of a happy child as neutral. So having the child could be bad, but couldn't be good. But, on the Tricky Expectation View, the whole gamble counts as neutral (in expectation) if the expected value for the child is positive. So on the Tricky Expectation View, it can be neutral to do something which could

be bad but couldn't be good.

Even if this argument is rejected, we could change the example so that the mother would have a *different* child in each of the four outcomes. This would mean that there was no one child who had a good prospect, which would mean that the Tricky Expectation View would not even apply. But it would still be intuitively clear that Option 2 is better than Option 1.

#### 4.2.1.3 Unfavorable cases: extinction

Strict Asymmetric Views have their least plausible implications in cases of extinction. This is especially relevant to the subject of this dissertation, since we are considering appealing to these views to lessen the importance of decreasing existential risk or producing other trajectory changes.

Consider this problematic case, for instance:

> *Mass Sterilization*: Some terrorists engineer a highly contagious, incurable virus, and they spread it throughout the world. This virus causes sterilization in all people that are infected, but causes no other health problems. Within 150 years, no humans exist.

Although all Asymmetric Person-Affecting Views can tell a story about why it would be wrong for these terrorists to disperse this virus, it seems that they cannot tell the whole story. They can appeal to the fact that many people alive had an interest in having children or perpetuating human civilization, they cannot appeal the fact that it is a simply a great loss, in itself, for human civilization to come to an end. This is easily brought out by considering a variant of the case:

> *Voluntary Extinction*: All people collectively decide not to have any children. No one is ever made upset, irritated, or otherwise negatively affected by the decision. In fact, everyone is made a little better off.

As (Temkin, 2000, 2008) points out, it would be bad if this happened, the benefits to present people notwithstanding.

On an Asymmetric Person-Affecting View, we must count the interests of "extra" people if they have bad lives, but not if they have good lives. This leads to another troubling conclusion:

> *Mostly Good or Extinction*: In one future, all but a few people have excellent lives. But a very small percentage of these people suffer from a painful disease that makes life not worth living. In the other future, no people exist.[9]

[9] I draw this case from Parfit's discussion of the Absurd Conclusion (Parfit, 1984). Holtug (2004, pp. 139-140) and Sikora (1978) make similar arguments about a similar example.

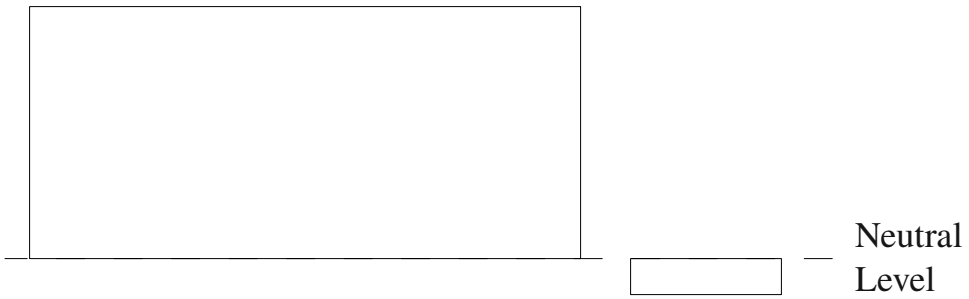


Figure 4.3: Mostly Good or Extinction (Version 1)

Intuitively, the first future is better than the second. But, given an Asymmetric Person-Affecting View, this is not true. On that view, all the good lives are ignored but the bad lives are not, and that makes existence worse than extinction. Ordinarily, we believe that there is a trade-off between bad lives and good lives. But on an Asymmetric View, we give no weight to the good lives, so the trade-off is not made properly.

We can extend our conclusions from Section 4.2.1.2 in disturbing directions when we think about sentient life in general. If we are merely thinking about how things go for future generations, and not thinking about ourselves, it seems that we should conclude that it would be best if sentient life in the universe came to an end. This would be so even if the chances of having excellent lives in the future were very high. Consider the following example:

> *The Anti-Biotic Explorers*: In the future, humans gain the ability to travel through space. They come across a planet that, as of now, has no sentient life. However, their best technical analysis suggests that if left alone, very happy sentient beings will soon evolve. Still, they admit that they could be wrong: life on this planet might turn out to involve some significant hardships for a few of these beings. For these beings, life might not be worth living. Since this planet has no valuable resources, and there is some small risk of having people with bad lives, the explorers decide to destroy this planet.

Intuitively, it was worse to destroy this planet rather than to let it be. However, for the reasons just given, Asymmetric Person-Affecting Views suggest otherwise.

Let's take stock of our conclusions about Asymmetric Person-Affecting Views so far. Both Person-Affecting and Unrestricted views can explain why we are typically not obligated to have

children. In terms of its benefits, the Asymmetric Person-Affecting View can explain why we need not procreate, but must not create people that will live in misery. As we saw in Sight or Paid Pregnancy, Asymmetric Person-Affecting Views can accommodate the judgment that we should make people happy instead of making happy people. Unfortunately, these views suggest that it is typically bad to have children, even in cases where one has a reasonable expectation that the one's child will have a happy life and it harms no one else. However, Asymmetric Person-Affecting Views lead to very implausible conclusions when we consider cases that highlight the value of large, flourishing civilizations. As we saw in Mass Sterilization and Voluntary Extinction, these views fail to accommodate compelling judgments about the importance of humanity's future. Since these views give weight to future people with bad lives but not future people with good lives, they also entail, implausibly, that a future that is good for most people, but bad for very few, is not worth having. Finally, as we saw in The Anti-Biotic Explorers, these views force us to be unreasonably averse to even a small risk that life will be bad for some people. This aversion is so great that we would prefer that no people exist, rather than risk that even one person have a life that is not worth living.

Not only are these consequences deeply implausible, they seem to be unintended side-effects of an attempt to accommodate a very narrow range of intuitive judgments. This gives us strong reason to reject this class of views. More reasons will emerge in the next section.

### 4.2.2 Moderate Person-Affecting Views

According to Moderate Person-Affecting Views, we cannot ignore the well-being of "extra" people, but their well-being counts for less. On the simplest version of this view, we add up the well-being of everyone who exists in an alternative to determine how good it would be, giving slightly less weight to the "extra" people. Much more sophisticated versions are possible as well.

These views need not have implausible consequences in Mass Sterilization, Voluntary Extinction, Mostly Good or Extinction, or The Anti-Biotic Explorers. Even at diminished weight, the interests of all the "extra" people in the future may be very significant because of the large number of people involved.

#### 4.2.2.1 Problems for any fixed weighting

Moderate Person-Affecting Views let us say some intuitive things about cases like Sight or Paid Pregnancy, but not all of them. On these views, it is good to create additional people, and it can

be better to create additional people even if it means that existing people will be worse off. Therefore, in some versions of these cases, Moderate Person-Affecting Views will have some implausible implications, at least if Unrestricted Views do.

Moderate Person-Affecting Views also face difficulties in Parfit's pregnancy cases:

> *The Medical Programmes:* There are two rare conditions, $J$ and $K$, which cannot be detected without special tests. If a pregnant woman has Condition $J$, this will cause the child she is carrying to have a certain handicap. A simple treatment would prevent this effect. If a woman has Condition $K$ when she conceives a child, this will cause this child to have the same particular handicap. Condition $K$ cannot be treated, but always disappears within two months. Suppose next that we have planned two medical programmes, but there are funds for only one; so one must be canceled. In the first programme, millions of women would be tested during pregnancy. Those found to have Condition $J$ would be treated. In the second programme, millions of women would be tested when they intend to try to become pregnant. Those found to have Condition $K$ would be warned to postpone conception for at least two months, after which this incurable condition will have disappeared. Suppose finally that we can predict that these two programmes would achieve results in as many cases. [Either one will decrease the total number of children with disabilities by 1000.] Parfit (1984, p. 367)

Suppose we fund Post-Conception Screening. The children of women with Condition $J$ would have existed regardless of what we did, but this is not true for the children of women with Condition $K$; this is because if we tell women with Condition $K$ about their condition, they will wait to conceive, and different children will exist. Therefore, the children who would have existed if we funded Pre-Conception Screening are "extra" people; because of this, if we adopt any kind of Person-Affecting View, funding Post-Conception Screening was better than funding Pre-Conception Screening. Intuitively though, it seems that the two programs are equally good.[10]

Some people claim that when they consider this case, they find that Post-Conception Screening is better for precisely the reasons stated above. If they believe this, we can ask how much better they think it is. Suppose, for instance, Post-Conception Screening only prevented half as many handicaps as Pre-Conception Screening. Unless we assign very significant weight to "extra" people, all Moderate

[10]If you assume that fetuses are not yet people, Presentism entails that both groups of children are "extra." Thus, this view implies that either program is equally good. However, according to Presentists, programs that prevent mothers from transmitting HIV to children that have already been born are better than programs that prevent mothers from transmitting HIV to their unborn children, even if the latter kind of program prevented significantly more instances of HIV. This is implausible for the same kinds of reasons that it is implausible to favor treating one group in *The Medical Programmes*.

Person-Affecting theorists must hold that Post-Conception screening would be preferable even in this case. It is hard to believe that Person-Affecting considerations could make Post-Conception Screening that much better than Pre-Conception Screening. To accommodate this judgment, we must place very significant weight on the interests of "extra" people (at least 50%).

To be fair, we should admit that Unrestricted Views will have their share of problems in variations of this case. Rather than funding Pre-Conception or Post-Conception treatment, defenders of Unrestricted Views must claim it is better to pay 1000 women to have healthy, non-blind children, as in Sight or Paid Pregnancy. But we have already acknowledged this problem.

Another case is quite problematic for Moderate Person-Affecting Views. Consider:

> *Disease Now or Disease Later*: A non-fatal disease will harm a large number of people. It will either do this now, or it will do it in the future. If the people are affected in the future, a greater number will be so affected; which future people exist will depend on our choice. Whatever we do, everyone will have a life that is worth living. (Doing it later will have no desirable compounding effects.)

According to Moderate Person-Affecting Views, both Symmetric and Asymmetric, it would be better to let future people face the disease. How much extra harm we are willing to tolerate will depend on our choice of weighting. Again, this puts pressure on us to make sure the weighting is fairly high. (The lower it is, the more "extra" people we will allow to suffer in order to protect people alive today.) The source of the problem here is that while it may be intuitive that it is better to help present people than to create additional happy people, it is not very plausible that it is less important to prevent harm to future people.

#### 4.2.2.2 Problems for different fixed weightings

Notice how the challenges facing Moderate Person-Affecting Views interact. To avoid implausible conclusions in The Medical Programmes and Disease Now or Disease Later?, the weight assigned to potential people must be reasonably high. To avoid implausible conclusions in Sight or Paid Pregnancy, the weight must be fairly low. Of course, this is unsurprising: having a very low weight for "extra" people is like accepting some kind of Strict Person-Affecting View, and accepting a very high weight is very much like accepting an Unrestricted View.

Often, in this kind of case, a moderate path will seem reasonable. But here, a moderate path seems to have little speaking in its favor. A moderate weighting (roughly 50%, say) would have implausible consequences in all of these cases. In The Medical Programmes, post-conception screening

would be about twice as good, and it would be better for diseases to strike in the future, even if they affect twice as many people as they would if they affected people now. Still, it would be good to create additional people, so creating happy people might be better than going around and curing blindness. (And of course, cutting the value of humanity's future in half will do little to undermine the view that the far future is overwhelmingly important.)

A fairly high weighting is also unattractive. It doesn't behave significantly better than an Unrestricted View in any respect. It has all of the implausible consequences in the case of Sight, or Extra People?. It avoids extreme implausibility in The Medical Programmes and Disease Now or Disease Later, but it finds differences between alternatives that are, intuitively, equally good. Very little, if anything, speaks in favor of this view.

That leaves us with the option of a low weighting. Even a very small weighting could explain what is wrong with all of the extinction cases from Section 4.2.1, so this option avoids the danger of repeating the problems of the Strict Person-Affecting View.

Of course this view will still have problems with cases like The Medical Programmes and Disease Now or Disease Later?. More troublingly, this view faces a sort of dilemma. There are two kinds of Moderate Person-Affecting Views: Symmetric Views and Asymmetric Views. Symmetric Views give diminished weight to the interests of "extra" people, whether their well-being is positive or negative. Asymmetric Views give diminished weight only to the interests of "extra" people with positive well-being. On Asymmetric Views, the interests of people with negative well-being are given full weight.

Given a low weighting, which is the alternative we are now discussing, Moderate Symmetric Views face a problem in the following kind of case. Suppose we must choose between creating the Wretched Child or allowing some person who is already alive to suffer from a less painful condition. Intuitively, it would be better to allow someone who already exists to suffer the less painful condition. However, if we give low weight to the interests of "extra" people (even if they have lives that are not worth living), then it would be better to create the child, even though he would suffer very much more than the person already exists would suffer. But that is very hard to believe. Even if we think that creating happy people is less good than making people happy, we cannot accept that making people miserable is much worse than making miserable people.

If we opt for a Moderate Asymmetric View with a low weighting, the challenges in Mostly Good or Extinction and The Risk-Averse Mother return to us. We must ask, "How much weight are we giving to the interests of the "extra" people with positive welfare?" Suppose we answer: 5% as much as we give to people who already exist (we must answer in this sort of way if we want to avoid the

pitfalls of moderate/high weightings). It is fairly hard to believe that a world like the one below would be bad:

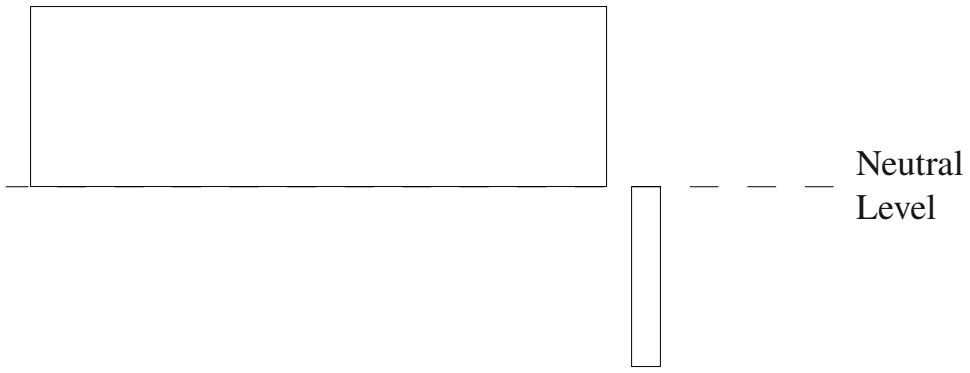


Figure 4.4: Mostly Good or Extinction (Version 2)

But it seems that it would be if we adopted this factor as our weighting for "extra" people.[11]

But this is hard to believe. It is challenging enough to create a society where most people are happy. If we were so successful that 95% of the population would have good lives, whereas 5% would have bad lives, how could we conclude that this would be *worse* than there being no future people? Some people might believe that though it would be good for us to have a future like this, this future would be bad if there had already been a lot of good lives. This intuition is best captured by some type of view according to which population has diminishing marginal value. I discuss these positions in the next chapter.

A variation on The Risk-Averse Mother is troubling as well. In this case, the mother chooses between two prospects for her child.

| | 95% | 5% |
|---|---|---|
| Option 1 | Rich and flourishing life (100 utility) | Bad life (-100 utility) |
| Option 2 | Neutral life | Neutral life |

With only a 5% weighting to future benefits, the expected value of Option 1 would be

$$0.95 \times 100 \times 0.05 + 0.05 \times -100 = -0.25$$

[11]This calculation assumes a simple modified-utilitarian weighing (modified to reflect the discounting). But that is generous to my opponent. If we adjusted for non-utilitarian factors such as priority/equality/maximin, that would only make creating the 95/5 world worse.

Thus, Option 1 would be worse than Option 2, on this view. But intuitively, that is wrong.

So, sticking to the idea of a low weighting, the dilemma is as follows. If we opt for a Moderate Symmetric View, we cannot accommodate the datum that making miserable people is as bad as making people miserable. If we opt for a Moderate Asymmetric View, the problems of Mostly Good or Extinction and The Risk-Averse Mother come back to us. Neither option is terribly attractive, so we have additional reason to reject these views.

## 4.3 Conclusion

So far, we have seen a variety of motivations and problems for Person-Affecting Views, both theoretical and intuitive. In this section, I will take stock of the situation and argue that, all things considered, Unrestricted Views fare better. They avoid the theoretical and intuitive problems of Person-Affecting Views, and their own problems are less significant. I'll end with some thoughts about where this leaves the rough future shaping argument.

### 4.3.1 Taking stock: costs and benefits of Person-Affecting Views

Let's begin with the the theoretical problems. The Person-Affecting Argument has fairly plausible premises, but a perfectly analogous argument entails that there is nothing wrong with creating a child that will live in misery. If anything is clear, it's clear that it would be bad to create a child that will live a life of unrelieved suffering. Therefore, this argument should be rejected. The second argument, which appealed to a distinction between victims and beneficiaries, was unpersuasive as well.

Person-Affecting Views face a variety of intuitive problems as well. Strict Asymmetric Views seem to drastically underestimate the importance of a good future for sentient life, as we saw in Mass Sterilization, Voluntary Extinction, Mostly Good or Extinction?, The Anti-Biotic Explorers. They also have implausible consequences in relatively ordinary cases of procreation, including The Risk-Averse Mother.

Moderate Person-Affecting Views can avoid most of these problems. On these views, it is good to create people with good lives, but the interests of "extra" people are given diminished weight. Depending on how much we diminish the weight of these people's interests, we face various problems. If we diminish it significantly, the view has implausible consequences in Parfit's case of The Medical Programmes, in Disease Now or Disease Later?, and in versions of Mostly Good or Extinction and The Risk-Averse Mother. If we diminish it only slightly, the view still has bad consequences in The

Medical Programmes and Disease Now or Disease Later? and doesn't behave too differently from Unrestricted Views in cases that defenders of Person-Affecting Views find problematic.

The case we considered which most favors Person-Affecting Views was Sight or Paid Pregnancy. It is intuitively plausible that it is better to benefit existing people than it is to create happy people, and this intuition is hard to capture without Person-Affecting Views.

### 4.3.2 Taking stock: costs and benefits of Unrestricted Views

How do Unrestricted Views fare in relation to all of these issues? For one thing, they avoid all of the theoretical and intuitive problems that are faced by Person-Affecting Views. This is a huge benefit. For another, they are very simple. One requires no complicated theory explaining whose interests shall be counted or what weight should be given to each. The interests of all people are treated the same, on this kind of view. (Of course, this doesn't mean that we can't count the interests of the worse off at a higher rate or that we can't care more about the virtuous. We simply mean here that we are not going to treat regular people differently from "extra" people.)

These views can also have the most important theoretical benefit, a benefit traditionally thought to go only to Person-Affecting Views. Though the Person-Affecting Argument fails, it is perfectly possible to accept an Unrestricted View and maintain the idea that morality is all about what is good or bad for individuals. In general, when we find some very natural-seeming idea and attempts to formalize it cause severe problems, that is good evidence that we have failed to appropriately formalize the idea. Rather than accepting the comparative Person-Affecting Restriction, those with Unrestricted Views could accept an absolute formulation:

> Personal Good Restriction: A possible world is bad only insofar as it is bad for particular people in that possible world, and an outcome is good only insofar as it is good for particular people in that possible world.

This kind of view can accommodate most of the theoretical applications of the Person-Affecting Restriction. This view can explain why it is not good to implement a government policy that benefits no one, and not good to stop people from doing things that are bad for no one. It justifies the use of Pareto principles in economic arguments. It explains why many people find it compelling to think that Mere Addition cannot make an alternative worse: it involves nothing that is bad for anyone. It also explains why, as many believe, there is nothing good about Leveling Down. Whether or not we ultimately want to accept all of this reasoning, we should admit that the Personal Good Restriction can work just as well as the Person-Affecting Restriction for these purposes. Insofar as

these are benefits of adopting the Person-Affecting Restriction, these are benefits of the Personal Good Restriction as well. Before moving on, we should note that the Personal Good Restriction is very general and abstract. It leaves open many possible views about how considerations for and against different actions should be weighed. It does not, for instance, immediately commit us to utilitarianism.

How does the Personal Good Restriction fare with respect to the intuitive motivation for accepting a Person-Affecting View? Clearly, some views that satisfy the Personal Good Restriction fare rather poorly. Those who accept the Personal Good Restriction must accept that it is good to have additional children and perhaps they must accept that it would be better for the wealthy philanthropist to pay people to have children rather than to cure the blind (at least in a cleaned up version of the case where having this children won't have very undesirable side effects). These things may be hard to believe, but I submit that they are not as hard to believe as the troubling consequences of Person-Affecting Views that have been described above.

Here is a table summarizing my findings across the different kinds of problematic cases:

| | The Happy Child | The Wretched Child | Obligation to have kids? | Sight or Paid Pregnancy | Repugnant Conclusion | Better to create happier people? | Bad to have kids? | Extinction Cases | The Risk-Averse Mother | The Medical Programmes | Disease Now or Disease Later? | Mostly Good or Extinction |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Strict Symmetric | ? | X | | | X | X | | X | X | X | X | |
| Strict Asymmetric | ? | | | | X | X | X | X | X | X | X | X |
| Moderate Symmetric (low weight) | ? | X* | | | ? | | | X | | X | X | |
| Moderate Asymmetric (low weight) | ? | | | | ? | | X* | | X* | X | X | X* |
| Unrestricted | ? | | | X | ? | | | | | | | |

X: The view faces problems with this case/principle

X*: The view faces problems with a version of this case

?: It isn't clear whether the view has intuitively implausible implications about this case

Table 4.2: Summary of findings regarding Person-Affecting Views

Obviously, with any table like this, or any selection of cases, it is possible to bias one's findings by only thinking about cases that favor one's preferred position. However, I find the different kinds of cases here to be representative of fairly natural classes of problems that one encounters when trying to develop any of the views listed above. In light of how this has turned out, and in terms of the

severity of the problems faced by these views, I think the Unrestricted View fares best of all the views that have been considered.

### 4.3.3 Where does this leave the value of shaping the far future?

In the future, we may adopt some other view rather than an Unrestricted View of the sort I have defended. This view might appeal to some of the positions I've considered in some contexts and other positions in other context. For each problem, there is a view that avoids it, so it would be possible to come up with some type of hybrid view that avoided all of the problems that I've discussed. What could we expect such a view to say about the value of ensuring that there are people in the far future? When we considered cases involving extinction and the far future, we found that Unrestricted Views had the most plausible implications. Therefore, we should expect that future theories would have similar implications about this issue.

Some may be tempted to hold onto some kind of Moderate Person-Affecting View, perhaps on the grounds that the cases that motivate the Unrestricted View are too far removed from everyday experience to be taken seriously. I reject this reasoning. It is important that we develop an adequate theory for evaluating the long-term prospects of human civilization, and this will often be simplest if we think about idealized cases first, rather than trying to handle more complicated cases first. But we would have to assign very low weight to "extra" people in order to undermine the view that shaping the far future is overwhelmingly important. In any case, there will be a vast expected number of potential people with good lives. So even if we accept a rather low weight (such as 1%) for our Moderate Person-Affecting View, shaping the far future will still be very important.

# Chapter 5

# Does Future Flourishing Have Diminishing Marginal Value?

## Introduction

Imagine the next 1 million years of human civilization go reasonably well in the respects that you think matter, about as well as things are going now, or perhaps a bit better. How important would it be that the following million years go equally well in these respects? And the million years after that?

Consider three kinds of answer:

1. The Period Independence answer: Equally as important in each such case.[1]
2. The Capped Model answer: After a while, it gets less and less important. Moreover, there is an upper limit to how much value you can get in this way.
3. The Diminishing Value answer: After a while, it gets less and less important. However, there is no upper limit to how much value you can get in this way.

Here is a graph illustrating the different styles of answer.

[1] Recall that Period Independence is the following claim:

> *Period Independence*: By and large, how well history goes as a whole is a function of how well things go during each period of history; when things go better during a period, that makes the history as a whole go better; when things go worse during a period, that makes history as a whole go worse; and the extent to which it makes history as a whole go better or worse is independent of what happens in other such periods.

I defended this claim in chapter 3.

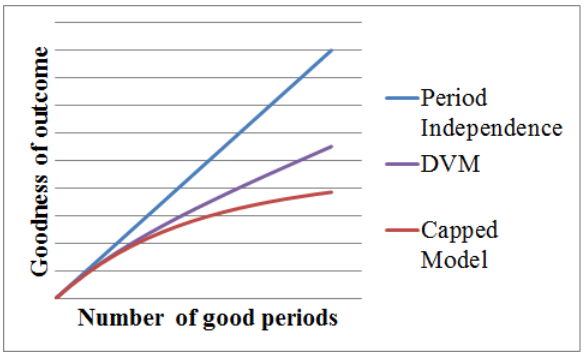


Figure 5.1: Period Independence, Diminishing Value Models, and Capped Models

On the x-axis, we have the number of periods, value on the y-axis. The Period Independence assumes a straight line, Capped Models assume a horizontal asymptote, and Diminishing Value Models assume decreasing slope but no horizontal asymptote.

## Narrowing our focus

As noted in chapter 3, many people will want to know more before they answer. People concerned about *peaks* (a perfectionist value) will want to know whether the very best aspects of our civilization will surpass the best of what has gone before. People concerned about *variety* will want to know whether people in the next period will be flourishing in new and different ways. People concerned about *troughs* will want to know whether the last period involved unprecedented levels of suffering, thinking that it will matter more if no record-breaking suffering happens in the next period. People who care about *shape* will want to know if things have been getting better than they used to be, caring whether we are on an upward trajectory or not. And people who care about *averages* and cross-period *equality* will want to know quite a bit about how things went in the distant past.

I said something about all of these ways of resisting Period Independence in chapter 3. I will not be taking account of these concerns in this chapter. Insofar as it is possible, I have chosen examples that leave things equal with respect to these concerns.

In the previous chapter, we discussed Person-Affecting Views. People with these views will want to have information about the temporal and modal profiles of the people in the future periods (Do any of them exist yet? Will they exist regardless of what we do? Etc.). I argued that we should reject these views, at least in the context of future generations. For the sake of argument, I will assume these views are wrong.

My argument for the overwhelming importance of shaping the far future assumes Period Independence. If we instead accept a Capped Model or Diminishing Value Model, this might affect our conclusions about the importance of the far future. It would be nice to know whether we could still reach the conclusion that shaping the far future is overwhelmingly important if we must reject Period Independence.

In this chapter, I identify the costs and benefits of the three kinds of model under discussion and explain what their implications are for determining the value of the far future. The main claims I will defend are:

1. Period Independence has the most plausible results overall.

2. Many of the most plausible versions of Capped Models and Diminishing Value Models do not undermine the view that shaping the far future is overwhelmingly important..

In the first section, I discuss the costs and benefits of Capped Models. In the second section, I discuss the costs and benefits of Diminishing Value Models. In the third section, I discuss how Capped Models and Diminishing Value Models bear on the Repugnant Conclusion and some related problems. In the last section, I summarize the main costs and benefits of both models in a table.

## 5.1 Capped Models and Period Independence

### 5.1.1 Temkin's Capped Model

There are three main features to Temkin's Capped Model (Temkin, 1997, 2012, ch. 10):

1. Pluralism about what matters

2. An upper limit to how much particular things could matter

3. Additivity across what matters

Let me briefly explain these elements with a couple of examples from Temkin (2012).[2]

When judges try to assess how good a gymnastics performance is, they use a model that incorporates all three of these features. The model is *pluralistic* because they care about how well the gymnast does in each of several different events, with its own standards of assessment (uneven bars, balance beam, floor, etc.). Since they give a score between 1 and 10 for each event, there is an

[2]Ranking systems with these properties are sometimes called "bounded multi-attribute utility functions" in economics.

*upper limit* to how much an excellent performance on any one event could matter. And the model is *additive*, since they determine an overall score by adding up the gymnast's score for each of the original events. This is why Temkin (1997) originally called this the "Gymnastics Model of Moral Ideals."

You might think we do something similar when we try to assess how virtuous a person is. In order to be very virtuous *all things considered*, it is not enough to superbly exemplify one or two of the virtues. There are multiple dimensions of virtuousness, and it matters how virtuous one is along multiple dimensions (this is the pluralism part). As Temkin (2012, p. 334) points out, even if Attila the Hun were very courageous and loved his mother a lot, we should not conclude, on those grounds, that he is virtuous. Though possessing any given virtue to a greater extent might make Attila more virtuous, there is an upper limit to how much his courage and love for his mother can contribute to his overall level of virtue. Since he is doing so badly in terms of the other virtues, no amount of courage and mother-loving can make him a good person overall. It is unclear that we will try to calculate a person's overall virtuousness by adding up his scores for all of the relevant virtues, but it is not implausible that this would give reasonable answers.

These three features of Capped Models are relatively independent. For my purposes, what matters most is the idea that there is an upper limit to the value of certain features of outcomes. I am less concerned about the model's pluralism, and even less concerned that we keep an additive version of the model. Someone might think that if you get rid of the pluralism, Capped Models no longer make sense. To illustrate this point, someone might argue as follows.

> If you applied Capped Models to a monistic theory like total utilitarianism, it is unclear what that would even mean. You might claim that there is an upper bound to the value of total utility, but you would rank outcomes the same way that you always did.

This argument assumes that total utilitarianism only provides an *ordinal* ranking of outcomes. We cannot express the claim that only total utility matters and that total utility has diminishing marginal value if we are only allowed to make ordinal comparisons over states of nature. But a Capped Model utilitarian could claim that the value of total utility is representable *cardinally*, and that it has diminishing marginal value. This would be a substantive difference between a Capped Model utilitarian and an ordinary, unbounded utilitarian. The difference between these views comes out most clearly when we consider how the two kinds of utilitarian will behave under risk. If they both rank prospects in terms of expected value, a Capped Model utilitarian will behave in a more risk averse fashion in cases with a large potential upside. I will say more about the relationship

between Capped Models and risk aversion in the next chapter.

### 5.1.2 How Capped Models challenge the case for the overwhelming importance of the far future

If we apply a Capped Model to the value of future periods, then we accept a view that is inconsistent with Period Independence. Since my argument for the overwhelming importance of the far future assumes Period Independence, this application of a Capped Model challenges my argument. Do Capped Models undermine all reasonable variations of my argument? No. That depends on whether averting an existential catastrophe could produce an amount of value that would take us a long way toward the cap. To see why, consider two possibilities:

1. If we reach the cap while humans (or our descendants) are still on Earth, or we start getting close to it, then that may significantly decrease the expected value of the future. For example, if things are as good as they can get regarding well-being after, say, 100 million years, then there might be very little (expected) value in ensuring that there are future generations after that. In that kind of case, we should be more concerned about the well-being of people alive today, since we have better methods of helping them and there is not yet much diminishing marginal value to saving their lives.

2. On the other hand, if the cap doesn't kick in until hundreds of trillions of years from now, then the cap may not significantly affect the value of the far future.

Though Temkin defends the use of Capped Models in some contexts and using Capped Models in this context could challenge the case for shaping the far future, it should be noted that Temkin (2012, ch. 10) rejects using Capped Models to argue that there is an upper limit to the value of additional future generations.

### 5.1.3 Considerations in favor of adopting a Capped Model across periods

Why might we adopt a Capped Model for comparing the value of different futures of the world? I will consider a few arguments in this subsection.

#### 5.1.3.1 From caps within periods to caps across periods

Much of the population ethics literature has focused on, well, populations. And it is natural to think about populations as existing in a given period, rather than across periods. And many people think

that, within periods, putting upper limits on the value of additional people is plausible. Consider, for instance, the following case:

> *The Double Population Policy*: If we implement the Double Population Policy, it will be possible to double the number of people living on the Earth. This, we stipulate, will not have any bad consequences with respect the use of scarce resources or environmental damage, it will not make the lives of the original people go any better or worse, and the lives of the new people will go equally as well as the lives of the old people.

For total utilitarians, implementing this policy would make things *twice as good* as they were before (or twice as bad, if one thinks the lives are, on the whole, bad). For many (and probably most) other people, it may seem natural to believe that the benefits of this policy would be much more limited. Some may even claim that there would be essentially no improvement if we enacted this policy.

Those who accept the latter judgment need not believe that the size of the total population is *irrelevant*. Some think that if there were not very many people, it would be very important to double the size of the world's population. Rather, these people seem to think that though adding additional people would be valuable in a context in which the population is small, adding them has relatively little value when we already have "enough" people. These people may even believe that there is an upper limit to the potential value of creating additional people.

I do not mean to rehearse the arguments in favor of adopting a view of this kind, or assess the plausibility of these arguments. What we say about these kinds of cases does not immediately say anything about the question that is presently of interest to us, which is: Is there an upper limit to the value of adding additional good periods of civilization?

However, there is a connection between what we say about adding additional good periods of civilization (and therefore additional people) and what we say about increasing the number of people living within a period. Someone might argue as follows:

(5.1) There is an upper bound to the value of ensuring the existence of additional people within periods.

(5.2) If there is an upper bound to the value of ensuring the existence of additional people within a period, then there is an upper bound to ensuring the existence of people in future periods.

(5.3) Therefore, there is an upper bound to the value of ensuring the existence of additional people in future periods.

I call this *Space/Time Symmetry Argument*. To my knowledge, no one has made this argument in exactly this form, though I suspect some philosophers would accept it. Let's take the premises in turn.

Why accept (5.2), and why call it a Space/Time Symmetry Argument? Imagine a two-dimensional space partitioned along intersecting perpendicular lines which form a grid. And imagine that valuable events could occur in these squares. How could we treat the two dimensions of space asymmetrically? We would do this if we claimed that there was an upper bound to the value of good events occurring within rows but not columns, or columns but not rows. This is analogous to how we would be treating time and space asymmetrically if we claimed that there were an upper bound to the value of additional people *within* periods, but not *across* periods. We could extend our 2D grid into a 3D grid by adding time as a third dimension. If we claimed that good events occurring within any temporal band had an upper limit to their potential value, but did not make an analogous claim about any events occurring within spatial bands, we would be treating space and time differently. Especially if we are skeptical of a deep metaphysical distinction between time and space, we might be skeptical of this asymmetry.

Such an asymmetry may also have strange practical implications. Suppose that there were an upper bound on the value of ensuring the existence of additional people within periods but not across periods and consider this case:

> *Sam's Delayed Birth*: We have a sperm and an egg which we can use to create Sam now, or in the future. If we create Sam now, his life will go somewhat better, and the timing of his existence will not affect anyone else's well-being. However, an enormous people are alive right now. In the future, fewer people will be alive.

In this case, we have reached the upper limit for the value of creating people *now*, which means that there would be very little value in creating Sam now, *in this period*. But, if we create Sam later, there will be fewer people in that period, so we will not have reached the upper limit for the value of creating additional people *in that period*. Our asymmetric treatment of creating lives in the current period and creating lives in future periods would support creating Sam in the future, though it is worse for Sam, better for no one, and has nothing to do with fairness, desert, rights, or any sort of deontological consideration. When discussing the Absurd Conclusion, Parfit (1984) claims that it would be absurd to think that we could make a situation significantly better or worse simply by changing around the order in which people are born, provided people would have equally good lives either way. It would be natural to make similar claims about examples like this.

#### 5.1.3.2 The Extra Colony, The Last Colony, and the Delayed Colony

In some cases, however, it may seem that we do treat space and time differently. Consider two hypothetical cases:

> *The Extra Colony*: In the next 1000 years, humans get the chance to colonize another planet. They know that if they succeed in colonizing this planet, then: (i) the new planet will sustain a population equal to the size of the population of the Earth, and this planet, like Earth, will sustain life for 1 billion years, (ii) these people's lives will probably go about as well the lives of the Earth people, (iii) life on both planets will last 1 billion years either way.
>
> *The Last Colony*: Human civilization has lasted for 1 billion years, but the increasing heat of the sun will soon destroy all life on Earth. Humans (or our non-human descendants) get the chance to colonize another planet, where civilization can continue. They know that if they succeed in colonizing this planet, then: (i) the new planet will sustain a population equal to the size of the population of the Earth, and this planet, like Earth, will sustain life for 1 billion years, (ii) these people's lives will probably go about as well the lives of the Earth people, (iii) there will not be a chance for the people on the new planet to colonize another planet.

Intuitively, it would be extremely important to colonize the extra planet in the second case, much more important than colonizing in the first case. Yet, we have supposed, there will be a similar number of people and a similar number of independent generations either way. Indeed, we could set the case up so that the very same people would exist either way. It might seem that we should explain the intuitive difference in terms of a fundamental difference between space and time. If we should treat time and space differently, then the Space/Time Symmetry Argument cannot threaten Period Independence.

However, there is reason to doubt that an asymmetry between time and space explains the intuitive difference between the Extra Colony and the Last Colony. To see this, consider the following example:

> *The Delayed Colony*: Just like the Extra Colony, but it will take our spaceship 1 billion years to reach the new planet that will be colonized. The colonists will be held in stasis and revived upon arrival. Because of the delay, once the colonists reach the planet, there will no longer be any people left on Earth.

If the asymmetry between space and time justified our different intuitions about The Extra Colony and The Last Colony, then colonization would have to be much better in The Delayed Colony than it is in The Extra Colony, but that is hard to believe.

Someone might insist that colonization is more important in The Delayed Colony than it is in The Extra Colony. This is intrinsically rather implausible, but further arguments may be offered against it. Suppose that we faced a choice between colonizing a distant planet, as in The Delayed Colony, and a closer planet, as in The Extra Colony. On the view in question, it would be much better to colonize the distant planet. We can ask these people:

> How much better do you think it would be? Since you think it is much better to colonize the distant planet, you presumably think that it would be reasonable to sacrifice the colonists' quality of life, at least to some extent, in order to travel to the distant planet. How much would you be willing to sacrifice? 10%? 20%? 50%?

Clearly, it would not be reasonable to colonize a more distant planet if that meant the inhabitants' lives would be significantly worse. But if we are only willing to sacrifice little, it is not plausible that we think there is a great difference between colonization in the two cases.

There is an additional problem with treating space and time asymmetrically, which comes from the special theory of relativity.

> *The Separated Worlds*: There are only two planets with life. These planets are outside of each other's light cones. On each planet, people live good lives. Relative to each of these planets' reference frames, the planets exist at the same time. But relative to the reference frame of some comet traveling at a great speed (relative to the reference frame of the planets), one planet is created and destroyed before the other is created.

If we treat space and time asymmetrically, we would have to claim that, relative to the reference frame of the planets, this outcome was not as good as it is relative to the reference frame of the comet. But this is very hard to believe. The value of this possible world should not be relative to any reference frame.[3]

Why might we think that colonization is more important in The Last Colony than in The Delayed Colony? Here are two potentially relevant factors. First, we might be concerned to have our culture, projects, science, and values develop for longer, independently of their contributions to our well-being. It is natural to think that these things would develop more in The Last Colony than in The

[3]I am grateful to Anders Sandberg and Carl Shulman for helping me to clearly formulate this objection.

Delayed Colony. Second, though we stipulated in the case that life would not be better during the last period of The Last Colony than it would be in the last period of The Delayed Colony, it is hard to imagine that this would happen. Perhaps, when we imagine this case, we do not properly acknowledge this stipulation. We could test this hypothesis by clearly stating that in The Last Colony, exactly the same things happen as happen on The Delayed Colony. When I make this stipulation, I no longer find it intuitive that one of these outcomes is better than the other.

Even if we accept the symmetry between space and time, the Space/Time Symmetry Argument can apply only minimal pressure on those who believe that there is no upper limit to the value of continuing civilization in a case like The Last Colony. It is very plausible that it matters greatly that civilization continues in cases like The Last Colony. It is less clear what we should think about cases like The Extra Colony. If we are forced to treat these cases alike, we should "inflate" the value of colonization in The Extra Colony and The Delayed Colony, rather than "deflating" the value of colonization in The Last Colony, as the Space/Time Symmetry Argument would recommend. Why? When we are resolving inconsistencies in our judgments, we should try to keep our confidence in the judgments about which we are most certain. The Space/Time Symmetry Argument requires us to give up a judgment about which we are quite certain so that we can hold onto a judgment about which we are less certain. The upshot of these considerations may be that space/time symmetry has to go, that we must "inflate" the importance of adding people on additional planets, or that we must appeal to some additional value such as the continuation of our culture, but these considerations count at best very weakly in favor of giving up our position that, other things being equal, it is of great importance that human extinction be delayed in cases like The Last Colony.

I am inclined to reject any fundamental normative asymmetry between space and time, and I therefore favor "inflating" our view about the importance of colonizing in The Extra Colony. Of course, I do not need this for my argument. If we reject this symmetry, then we can simply reject the Space/Time Symmetry Argument and ignore any plausibility it might have lent to Capped Models.

It may be worth noting that there may be some room in the position I favor for doing something that seems rather like treating time and space asymmetrically. One kind of rationale for accepting Period Independence, apart from accepting separability across different lives, is that there may be a special kind of value to closely integrated societies, but this value might diminish as different stages of civilization become less and less "connected" with each other. If one accepts this view, one can claim that doubling the population of a given civilization has relatively little value, provided one does so in a way that does not create additional societies (as one does in the colonization cases). Someone who accepts this view might claim that while doubling the population of the people in all civilizations

does little good, spawning new, separate civilizations (as in all of the colonization cases) is of great importance. This kind of position would not behave exactly the same as accepting an upper limit to the value of additional lives within periods. This position would, for instance, claim that colonizing in The Extra Colony was much more important than implementing the Double Population Policy. This position would also see no difference between The Extra Colony, The Delayed Colony, and The Last Colony.

#### 5.1.3.3 The Last Colony and The Very Last Colony

Those who wish to use caps across periods might conclude from the previous section that caps within periods work differently than caps across periods, or they may not. We might test the cross-period Capped Model more directly by constructing a case where a Capped Model had better have kicked in, on pain of near-total irrelevance. Compare The Last Colony with this example:

> *The Very Last Colony*: Convinced of the importance of preserving future generations, we take great precautions to protect the far future. Our descendants succeed in colonizing a large portion of the galaxy. It becomes relatively clear that our descendants will last for a very long time, about 100 trillion years, until the last stars burn out. At that point, there will be nothing of value left in the accessible part of the Universe. It comes to our attention that there is a chance to colonize one final place, just as in The Last Colony, before civilization comes to an end. For this billion years, these will be the only people in the accessible part of the Universe. During this period, things will go exactly as well as they went in The Last Colony.

In which case is colonization more important, The Last Colony or The Very Last Colony?

According to Period Independence, colonization is equally as important in each case. Intuitively though, it is more plausible to claim that colonization is much more important in The Last Colony. The thought seems to be that since 100 trillion years is so much more than 1 billion years, additional flourishing has less value if our descendants have survived for so long. If we accept a Capped Model we can, and probably should, try to accommodate the judgment that colonization is less important in the case of The Very Last Colony.

The defender of Period Independence could try to attribute this intuitive judgment to the *proportional reasoning fallacy*, which we discussed in chapter 2. We often look at proportions (as a heuristic) in cases where only *absolute* differences are much more important. As discussed in chapter 2, we use misguided proportional reasoning in some cases where many lives are at stake. Fetherston-

haugh et al. (1997) found that participants significantly preferred saving a fixed number of lives in a refugee camp when the *proportion* of lives saved was greater. Recall that, describing the participants' hypothetical choice, they write:

> There were two Rwandan refugee programs, each proposing to provide enough clean water to save the lives of 4,500 refugees suffering from cholera in neighboring Zaire. The Rwandan programs differed only in the size of the refugee camps where the water would be distributed; one program proposed to offer water to a camp of 250,000 refugees and the other proposed to offer it to a camp of 11,000.

Participants significantly preferred the second program. In another study, Slovic (2007) found that people were willing to pay significantly more for a program of the second kind.

It is as if the people in these studies think that additional lives matter less when the proportion is smaller. Similarly, we might claim, people think that creating another colony matters less when the proportion by which civilization is extended is smaller. (In The Last Colony, the duration of our existence doubles. In The Very Last Colony, it increases by a factor of $10^{-5}$.) More tests would be required to determine whether this bias is responsible for the relevant judgment, but the possibility of the proportional reasoning fallacy influencing this judgment should arouse our suspicion.

### 5.1.4 Difficulties with employing Capped Models across periods

#### 5.1.4.1 The Very Last Colony, The Even Greater Future, and the appeal to potential

Although using a Capped Model across periods might explain the judgment that colonization is less important in The Very Last Colony than The Last Colony, it does so at the risk of showing too little concern about the potential end of civilization.

Consider, for instance, what Capped Models have to say about The Very Last Colony. Even if we agree that colonization is more important in The Last Colony than in The Very Last Colony, we shouldn't (and don't) think that it is essentially unimportant whether colonization takes place in The Very Last Colony. But, if we thought there were an upper limit to the value of additional good periods, we would have to believe that it would not matter very much whether we succeeded in colonizing this planet. Or, if we don't believe this in that case, we would have to believe this in a case where there were vastly more periods of happy people. And even that is very hard to believe.

There is a variation on The Very Last Colony that has implications for Capped Models which are even harder to accept:

> *The Even Greater Future*: After our descendants survive for 100 trillion years, they discover that physics is rather different from how we supposed. It is possible for them to survive and flourish for another $10^{80}$ years rather than just another billion years.

On pain of near-total irrelevance, Capped Models imply that surviving and flourishing for another $10^{80}$ years has essentially no value. Why? If it we haven't reached the cap in 100 trillion years, the cap is unlikely to be relevant for any decisions we make. But if we have reached the cap by that time, it would be inconsequential whether our descendants survived for another $10^{80}$ years. And that conclusion is very hard to believe.

**The appeal to potential:** Those who defend a Capped Model in this context may reply by altering their views so that what counts as "near the cap" depends on how much *potential* there is for creating desirable outcomes. The idea is that we could have a very high cap in cases where it is possible for humanity to flourish for a very long time, and a lower cap in cases where it is possible to flourish for a shorter duration. This could help explain why it would be very important to colonize The Last Colony, but less important if we fail to colonize in The Very Last Colony, and it could also explain why it is very important to colonize in The Even Greater Future. How? In The Last Colony, colonizing would help humanity achieve a much greater proportion of its potential (50% vs. 100%). In The Very Last Colony, colonizing would only help humanity achieve a slightly greater proportion of its potential. And in The Even Greater Future, colonizing would give humanity access to the vast majority of its potential.

An obvious challenge to the appeal to potential is that there may be cases where there is no upper bound to how long civilization could flourish. We could imagine a situation in which someone could choose any number, and civilization would flourish for that many years. In this situation, it may seem that there is no way to apply the appeal to potential. However, it may be natural to say that in this case, since our potential is unbounded, there is no cap. This would be a departure from a Capped Model, but it would be a natural departure.

Though the appeal to potential avoids that problem it produces additional problems which make it unclear whether it is a genuine advance on the standard Capped Models. If we make the appeal to potential, the value of an outcome depends on which alternatives are available. For example, in The Very Last Colony, our options are:

(5.4) Survive and flourish for a total of 100 trillion years

(5.5) Survive and flourish for 100 trillion years, plus an additional billion years

Imagine a variant of The Even Greater Future, where there is a third option as well:

(5.6) Survive and flourish for 100 trillion years, plus an additional $10^{80}$ years

Now consider two possible worlds. In World 1, our options are (5.4) and (5.5), and we choose the former. In World 2, our options are all three, and we choose the first option. According to the appeal to potential, World 1 is significantly better than World 2. But that is hard to believe. World 2 may be more of a disappointment, and the decision-maker may be more blameworthy, but these facts could not make World 2 much better than World 1. More disturbingly, consider World 3. In World 3, we have all three options available, and we choose to survive and flourish for 100 trillion years, plus an additional billion years. According to the appeal to potential, World 3 is better than World 1. But this seems clearly wrong; by far the most relevant difference between these possibilities is that World 3 involves an additional billion years of flourishing, and that makes World 3 better than World 1 regardless of differences in potential. In light of these difficulties, I suggest that the appeal to potential is more trouble than it is worth.

#### 5.1.4.2 A risk-based objection

In other cases, considerations of risk might force Capped Models to positively advise *against* colonizing. To see this, suppose Capped Models apply across periods. Then we can make an argument against Capped Models that is analogous to an argument I made against Person-Affecting Views in the previous chapter.

(5.7) In the case of The Very Last Colony, we have reached the upper limit of value from additional good periods.

(5.8) Therefore, there is no potential upside to colonizing the last planet.

(5.9) However, there would be a potential downside. If we colonize the planet, there will always be some chance that things go badly—even if the colonists would have remarkably excellent prospects. Perhaps a completely unforeseen plague would make life miserable for the people there, for example.

(5.10) But, if some course of action has no potential upside and it has a potential downside, then choosing it is bad (in expectation)

(5.11) Therefore, it would be bad (in expectation) to colonize the last planet.

Since this conclusion is very implausible, we should reject a premise that led us to it. It seems more plausible to claim that Capped Models apply within periods, but not across periods, or never apply. Other replies, however, are possible. These replies have the same structure as the replies to the analogous arguments against Person-Affecting Views in the previous chapter. And these replies fail for analogous reasons.

First, someone who accepts a Capped Model across periods might reject (5.9), the assumption that there would be a significant downside if things went badly in the last period. Though it would be most natural to claim that creating additional periods where things go very badly would be very bad, this is not *entailed* by Capped Models. We could aggregate the value of all the good and bad periods together, letting the bad weigh against the good, and place an upper limit on the value of the whole (first version). Or, we could aggregate the value of all the good periods and place an upper limit on it, aggregate the value of all the bad periods and place an upper limit on that, and then let the results weigh against each other (second version).[4] If we go the first way, there would not be a significant downside to having a future period filled with great suffering for the same reason that having a future period filled with great happiness would not be a significant upside. Either way, we are already near the upper limit of how good things could be. However, filling any future period of the Universe with great suffering would make things significantly worse. Absent extremely powerful reason to give up this belief, we should hold onto it. Therefore, we should reject the first version of this reply.[5]

Second, someone might resist (5.10), claiming that some actions can be be good or neutral (in expectation), though they have no potential upside and have a potential downside. The most plausible way to do this would be to claim that when choosing uncertain prospects over future periods, it is best to maximize the expected quality of the future periods, but that the value of outcomes with very high expected quality should be discounted, where the strength of the discounting

[4]Here are two ways to spell out this thought formally (though others are possible). We give each period $p_i$ a score $v(p_i)$, representing how bad or good it is. We want to use this information to determine the value of the whole, $g(w)$. On one type of Capped Model, we do this by adding up all the $v(p_i)$ and then hitting $v(p_i)$ with a dampening function $f$ that has an upper bound. For example, we might use $f(x) = \arctan x$. So, on this version,

$$g(w) = f(\sum_i v(p_i))$$

This corresponds to treating quality of individual period as an individual ideal in a Capped Model.

On the second version, we determine $v(w)$ by dividing the periods into the good periods $G$ and the bad periods $B$, adding up the total good and hitting it with $f$, adding up the total bad and hitting it with $f$, and then calculating the difference. On this version, we might have

$$g(w) = f(\sum_{i \in G} v(p_i)) + f(\sum_{i \in B} v(p_i)).$$

[5]These points rely on an obvious debt to the discussion of the Absurd Conclusion in Parfit (1984).

increases if things went well in the past.[6] If we accept this model, we can deny that it would be bad (in expectation) to colonize in The Very Last Colony, since the expected quality of that future period is high. Let's call this the *Tricky Expectation View**, to distinguish it from the analogous view discussed in the previous chapter.

However, the Tricky Expectation View* is hard to accept because it is not consistent with the principle that it is bad (in expectation) to do things which could be bad and couldn't be good. After all, if life goes badly for people in The Very Last Colony, defenders of the Tricky Expectation View* will count this as a bad result, though they will count the production of good future as neutral. So colonizing in The Very Last Colony could be bad but couldn't be good. But, on the Tricky Expectation View*, colonizing The Very Last Colony counts as neutral in expectation. So on the Tricky Expectation View*, it can be neutral (in expectation) to do something which could be bad but couldn't be good.

#### 5.1.4.3 More on Our Surprising History and the appeal to a restricted scope of concern

Using Capped Models across periods also has implausible consequences in cases where we learn that the past, or remote regions of space, are different than we had previously believed. On such models, learning about these things can teach us that our fate matters much less than we previously supposed, and these consequences are hard to accept.

To bring this out in more concrete terms, consider again an example from chapter 3.

> *Asteroid Analysis*: World leaders hire experts to do a cost-benefit analysis and determine whether it is worth it to fund an Asteroid Deflection System. Thinking mostly of the interests of future generations, the leaders decide that it would be well worth it.

And then consider the following endings:

> *Our Surprising History*: After the analysis has been done, some scientists discover that life was planted on Earth by other people who now live in an inaccessible region of spacetime. In the past, there were a lot of them, and they had really great lives. Upon

[6] The simplest formal model of this that I can think of has the following form:

$$g(A) = \begin{cases} f(\sum_{i \in G} v(p_i) + EV(p)) + f(\sum_{i \in B} v(p)) & \text{if } EV(p) \geq 0 \\ f(\sum_{i \in G} v(p_i)) + f(\sum_{i \in B} v(p_i) + EV(p)) & \text{if } EV(p) < 0 \end{cases}$$

where $A$ is the act we are choosing, $p_1$ to $p_n$ are the periods in the past, $p$ is the future period that follows from choosing $A$, $v(p_k)$ is the quality of period $k$, $EV(p)$ is the expected quality of period $p$, $f$ is a dampening function that ensures that there is an upper limit on the value of choosing our action, $G$ is the set of good periods, and $B$ is the set of bad periods.

learning this, world leaders decide that since there has already been a lot of value in the universe, it is much less important that they build this device than they previously thought.

*Surprising Cosmology*: After the analysis has been done, some scientists find conclusive proof that we live in a Big Universe. The Universe is actually infinitely large. Based on some firm theoretical calculations, they show that somewhere in the universe, approximately $10^{1000}$ light years away, there are many, many civilizations full of achievement, happiness, and justice. Furthermore, this is almost certain to be true for the indefinite future. Upon learning this, world leaders decide that since there is already a lot of value in the universe, it is much less important that they build this device than they previously thought.

Many people find that, intuitively, it could not be less important to build the Asteroid Deflection System after learning these facts. The rough thought is that, typically, information about what happens in very distant regions of spacetime is irrelevant to how much good our actions will accomplish. Some others find that, intuitively, it could be less important to build the Asteroid Deflection System because we now know that we are not the only intelligent life. But very few would be inclined to say that learning about these facts would make it much, much less important to build the Asteroid Deflection System.

Yet, if we use Capped Models across space or across time, that's what we have to say. When we learn about Our Surprising History, we learn that there have been many good periods in the past. As was the case in The Very Last Colony, we will have already reached our limit on the value of ensuring that things go well in future periods. So it could easily be that before learning about our past, it would be very good to build the Asteroid Deflection System, but that afterward, it would matter little. Similar remarks apply to the case of Surprising Cosmology if we reject the asymmetry between space and time.

I should add that the Big Universe issue is not idle speculation. The dominant (but not necessarily consensus) view among astrophysicists is that we do live in a Big Universe (Bostrom, 2011; Knobe et al., 2006). Moreover, learning *whether* we live in a Big Universe could have very significant information value, given a Capped Model across periods. On this model, it would make good sense for altruists to spend significant time and effort trying to learn about whether we do live in a Big Universe, since it could have such significant policy implications. These consequences are hard to accept.

**Other unintended consequences:** People sometimes try to come up with Capped Models that will make all the same comparative value claims in cases with a small population as in cases with a large population, when comparing decisions that would only affect a small, disconnected part of the population. They want models that allow them to say that the evaluative status of cheating on your taxes, visiting your grandmother, or being a vegetarian is independent of whether or not you live in a Big Universe, though it matters less whether you do these things if you live in a Big Universe. The key is to claim that the value of all your options scale down uniformly as the population increases. But this has some unintended consequences, as the asteroid example shows.

It has other unintended consequences as well. To help illustrate an unintended consequence, first consider a view with agent-centered prerogatives, as in Scheffler (1994). On this kind of view, it is permissible to trade off 1 unit of your well-being against $n$ units of general goodness, though different versions of this view will claim that different kinds of trade-offs are permissible. Many people who want to acknowledge *some* duty to promote the general welfare but wish to place limits on how extensive this duty can be in ordinary cases, will be attracted to some version of Scheffler's position.

With caps across periods, as the number of good periods in the Universe increases, the value of your opportunities for doing good will decrease. Together with agent-centered prerogatives, this will imply that as the universe gets larger, you will be permitted to trade off your own interests against the interests of others differently than you otherwise would have in a smaller universe. This is because the increasing size of the universe makes your actions affect the overall value less (on a cross-period Capped Model), but does not affect how much your actions affects your own well-being. Therefore, it could turn out that whereas it would have been wrong to ignore the poor if you didn't live in a Big Universe, it is now permissible because you do live in a Big Universe and those reasons are much less weighty.

This problem might be avoided if we tried to argue that the reasons to look after one's own interests somehow scale down when one lives in a Big Universe. But that seems implausible and fairly ad hoc. Independently of trying to avoid this problem, few of us would be inclined to accept a view according to which our lives go worse simply because we live in a Big Universe. Just intuitively speaking, if someone learned that he lived in a Big Universe and then concluded that he now had less reason to live than he had before, this would seem very unreasonable. Though some people can work themselves into existential crises when they reflect on the fact that they are tiny specks in a vast universe, on reflection, this reaction seems unjustified (Nagel, 1971, pp. 716-718). And even if it were reasonable, few of these people would, upon learning that they live in a *infinitely*

*large universe*, decide that their lives matter *even less*. For my part, I am inclined to think that the value my life has for me—the value of my network of personal relationships and projects, goals and ambitions, my happiness, achievements and so forth—is thoroughly independent of whether or not I live in a Big Universe. In response to the problem I am describing, it will not be easy to claim that the personal value that our lives have for us diminishes depending on factors so removed from the parts of our lives that matter to us.

The moral here is not exactly that Capped Models across periods have a severe problem on the grounds that they interact with agent-centered prerogatives in a problematic way. Rather, the moral is that if we adopt a Capped Model across periods, we open ourselves up to difficulties that we may not have imagined.

**Restricting our scope of concern:** One natural way to describe the difficulties we face when considering Our Surprising History and Surprising Cosmology is to say that what happens in distant regions of space and time cannot be relevant to what it would be best to do, at least what happens there is beyond our control. We might say that these things are outside our scope of concern, or irrelevant for the purposes of choosing the action that would have the best consequences.

One way to make this thought precise is to consider the structure of a theory of value. Many theories of the value of outcomes have at least the following three parts:

1. Domain: A domain of objects, properties, and relations that matter
2. Value assignment rule: A rule for saying how good an outcome is if those objects, properties, and relations are in certain conditions
3. Scope of concern: A rule that says which of the objects in the domain are relevant for assigning value to outcomes, or for making goodness-motivated decisions

For classical utilitarians, the domain is the set of sentient beings and how happy they are, the value assignment rule is to add up the total happiness, and the scope of concern is unlimited: everyone's interests should be accounted for when promoting classical utilitarian values. We could also imagine a variation on this view according to which the scope of concern was more limited. For instance, the scope of concern could be limited to future events or events that are under the decision-maker's control. These changes would not yield different recommendations in ordinary cases, but it would change the structure of the theory.

In other cases, limiting the scope of concern could be decision-relevant. Person-Affecting Views can be thought of as theories where we keep the classical utilitarian value assignment rule and domain, but limit the scope of concern. For instance, presentists, who hold that only the well-being of people that presently exist matters for assigning value to outcomes, have a scope of concern limited to the people that presently exist. To take a more relevant example, on a Capped Model, it matters whether our scope of concern is universal or limited to events that are under our control. In the latter case, we can ignore the information we gain in Our Surprising History and Surprising Cosmology, which would give us intuitively correct results.

However, restricting our scope of concern introduces complications. Consider what happens if we restrict our scope of concern to events under our causal control. Imagine that in World 1, there is an enormous amount of suffering outside of our light cone, but things go very well inside of our light cone. In World 2, the reverse happens: lots of suffering inside our light cone, lots of good things outside of it. Suppose, as seems likely, there is a lot more stuff outside of our light cone than inside of it. In that case, World 2 would be better. But if we restrict our scope of concern to our light cone, we, implausibly, get the opposite conclusion.

Rather than claiming that what happens outside of the scope of concern cannot make outcomes better or worse, it would be more plausible for theories with a restricted scope of concern to claim that agents making goodness-motivated choices should only consider the effects of their actions on outcomes inside the scope of concern. The main problem with this idea is that, intuitively, the right way to make goodness-motivated choices is to *do as much good as possible*, and it is strange if this is not calculated by (i) considering all possible actions, (ii) calculating, for each action, how good it would be if that action were taken, and (iii) selecting the action that results in the best consequences. At least, that seems like the right way to do it when there are no relevant personal or deontological considerations at stake. But perhaps there could be some other way of understanding how it is proper to make goodness-motivated choices. For instance, perhaps the right way to make a goodness-motivated choice is to select the action that would be best serve one's reasonable altruistic concerns. Many variations are possible, and I won't try to explore them here.

To summarize, restricting our scope of concern can avoid implausible implications in Our Surprising History and Surprising Cosmology, but it does so at the cost of introducing other implausible implications or forcing a revision in ordinary ways of thinking about how it is best to make goodness-motivated choices.

### 5.1.5 The Aliens Objection

Some philosophers object to the rough future-shaping argument on the grounds that there is probably a lot of intelligent life in the universe, and that if there is a lot of intelligent life in the universe, then reducing existential risk is not very important from an impersonal perspective. My best attempt to reconstruct this objection is as follows:

1. During each period that humanity could be around, there will be many aliens in other parts of the universe.

2. During each period, there is an upper limit to how much better that period can get by creating additional sentient beings. (I.e., there are caps within periods.)

3. If there are many aliens during some period and there is an upper limit to how much value there can be due to the number of sentient beings that exist during that period, we are very close to the upper limit to how much value there could be during a period due to increasing the number of sentient beings that exist during that period.

4. Therefore, for each period that humanity could be around, we are very close to the upper limit to how much value there could be during that period due to increasing the number of sentient beings that exist during that period.

5. If, for each period that humanity could be around, we are very close to the upper limit to how much value there could be during that period due to increasing the number of sentient beings that exist during that period, then there is very little impersonal value in ensuring the existence of future generations.

6. Therefore, there is very little impersonal value in ensuring the existence of future generations.

Call that the *Aliens Objection*. In essence, the argument is claiming that because there are probably many aliens at all future times our descendants might exist, we're already effectively "maxed out" on the value of creating additional sentient life, and, because of this, there is very little impersonal value in ensuring the existence of future generations of humans.

It's very unclear whether the fifth premise is true because if the far future goes very well, then even if the impersonal value of additional lives is "maxed out," there could be value to ensuring the existence of future generations for other reasons. But I did not rely on these other reasons when making the rough future-shaping argument, so I will ignore this issue for the purposes of this discussion.

We should distinguish this version of the Aliens Objection from a more practical version which could be highly relevant, if it were true. On the practical version of the Aliens Objection, if we don't create an excellent future, other intelligent life forms will use the resources we would have used to create a future that is about equally as good. If this were true, then the opportunity cost of human extinction would be much less. But there is no reason to believe the key empirical claim of the practical version of the Aliens Objection because there is no reason to believe that there are any aliens nearby. On the version of the Aliens Objection I am discussing, the argument only assumes that there are many aliens somewhere very, very far away.

For the sake of argument, I'll grant the empirical claim about the existence of aliens. Due to the potentially enormous size of the universe, this claim is worth taking seriously.

I find this argument unconvincing because the key normative premise–premise two–is not clearly true, the conclusion is counterintuitive, and accepting the argument would have counterintuitive implications. On the first point, it's not clear whether premise two is true because, as we saw earlier in this section, claiming that there are caps across periods has counterintuitive implications, and claiming that there are caps within periods but not across periods has other counterintuitive implications. It is by no means clear that rejecting premise two is more implausible than accepting these counterintuitive implications. On the second point, as we saw in extinction cases discussed in the previous chapter—such as Mass Sterilization and Voluntary Extinction—there is intuitively a great deal of value in ensuring the existence of future generations, and it isn't just person-affecting value. If we consider the possibility that there is a lot of other intelligent life in the universe, it does not change our conclusion that there is a great deal of impersonal value in future generations. On the third point, if the objection holds, then, we're already effectively "maxed out" on impersonal value. And if we're already effectively "maxed out" on impersonal value, but not on person-affecting value, then strict person-affecting views are correct for practical purposes. But we saw in the previous chapter, strict person-affecting views have many counterintuitive practical implications, and these implications hardly become more plausible if we assume that there are many aliens. Moreover, many surprisingly ordinary actions, such as having children or distributing condoms in Africa, have different moral characteristics than they would if there weren't many aliens. Of course, this is highly implausible. For all these reasons, the Aliens Objection seems to give us little reason to resist the rough future-shaping argument.

## 5.2 Diminishing Value Models

In this section, I explain the Diminishing Value answer in greater detail, noting how it challenges Period Independence and the case for the overwhelming importance of shaping the far future. Next, I discuss some challenges that this model shares with employing Capped Models across periods, some advantages that this model shares with Capped Models, and some challenges that this model avoids.

### 5.2.1 How Diminishing Value Models challenge the case for the overwhelming importance of shaping the far future

On *Diminishing Value Models*, it matters less and less that additional periods go well. But there is no upper limit to the value of adding additional good periods. These models entail that the value of creating additional periods depends on what has happened in very distant periods. So, in order to know how important it would be that things go well in the next period, we would need to know how well things were going in the distant past. So, if this view were true, Period Independence would be false.

Could the argument for the overwhelming importance of the far future succeed if we adopted a Diminishing Value Model? Yes, but it depends on a couple of additional features of Diminishing Value Models: just how "diminished" value can get, and how many good periods we need for the diminishing value to start having significant effects.

#### The final rate of diminished value

We can partially characterize a Diminishing Value Model by asking what its *final rate of diminished value* is. This is a proportional measure of how much less value an additional good period provides in the limit, when we have already had a lot of good periods, and how much value an additional good period provides in the beginning. On a Capped Model, the final rate of diminished value is 0%.[7]However, the final rate of diminished value could be zero without having this model collapse into a Capped Model. This would mean that as the number of good periods became larger and larger, the value of an additional period would approach zero, but there would be no upper limit to

[7]Formally, the final rate of diminished value is $f$ when, for equally good periods $p_1, p_2, \ldots$, and the empty period $p_0$,

$$f = \lim_{n \to \infty} \frac{g(p_0 + p_1 + \cdots + p_n) - g(p_0 + p_1 + \cdots + p_{n-1})}{g(p_0 + p_1) - g(p_0)}$$

Within the scope of $g$, '$(p_a + p_b)$' denotes the world that is the result of following period $p_a$ with period $p_b$.

the amount of value that could be achieved by creating additional good periods.[8]

It is consistent, but rather strange, to have a theory that has no final rate of diminished value. This could happen if the value per additional period sometimes increased and sometimes decreased, no matter how many periods there already were. But it is plausible that if some Diminishing Value Model is correct, the value per additional period does not increase, at least not when there have already been a very large number of periods. Because of this, plausible versions of Diminishing Value Models will have a final rate of diminished value.

A Diminishing Value Model will significantly undermine the case for the overwhelming importance of shaping the far future only if the final rate of diminished value is very low. If the final diminished marginal value rate is low (such as 10%, say), then the case for the overwhelming importance of shaping the far future will not be significantly affected. If the final rate is close to zero, then the argument could easily fail, provided two other conditions are met, which we'll discuss below.

**How quickly does the diminishing marginal value set in?**

If we don't get a significant amount of diminishing marginal value for an extremely long time, then Diminishing Value Models would not affect my position at all. Therefore, it is not enough that the final rate of diminished value is very high, it also has to be true that we reach that rate (or some other rate of very significant diminishing value) fairly early on in humanity's existence.

On the other hand, if diminishing value sets in very quickly, that may *support* the case for the overwhelming importance of shaping the far future. To see this, suppose we believed that the final rate of diminishing value were 1%, and that it had already been reached, either because we've already had enough people on Earth or because we think we live in a Big Universe. In that case, making this period go well would have 99% less value than we previously thought. Since the *ratio* between all of the options would be unaffected, our ranking of options in terms of impersonal value would be completely unaffected (though, as pointed out in Section 5.1.4.3, it might change the all-things-considered ranking of options where more than axiological considerations are relevant).

Therefore, if diminishing value considerations are going to rebut the case for the overwhelming importance of shaping the far future, they must do so by claiming that future periods have diminishing marginal value in such a way that (i) we cannot be confident that the final rate of diminished value will be (or has been) reached, and (ii) the rate of diminished value must become small relatively soon (on a cosmic scale). If these conditions obtain, then our conclusions about the importance of

[8] If this sounds mathematically inconsistent, consider that the function $f(x) = \sqrt{x}$ has no upper limit, but its slope approaches zero as $x$ gets very large.

decreasing existential risk will be sensitive to the some of the key probability estimates about our future presented in chapter 3, together with information about the cost of decreasing the existential risk in question.

### 5.2.2 Virtues shared with Capped Models

Now that we know what is at stake with Diminishing Value Models, we can compare their plausibility with Capped Models. I'll begin by thinking about what virtues both kinds of model possess, and then turn to how they compare with respect to the challenges.

First, as with Capped Models, Diminishing Value Models allow us to preserve the symmetry between time and space, while still maintaining that creating additional people in a given period has diminishing marginal value. But, as with Capped Models, a potential cost of preserving this symmetry is to claim that colonizing would be equally important in The Last Colony and The Extra Colony, which is somewhat counterintuitive, at least at first.

Second, like Capped Models, Diminishing Value Models can justify the judgment that colonizing is significantly more important in The Last Colony than in The Very Last Colony, provided one does not approach the final rate of diminished value until rather late in the game.

### 5.2.3 Challenges shared/avoided

Next, we turn to some challenges that both Capped Models and Diminishing Value Models face, and some challenges Diminishing Value Models can avoid.

First, Diminishing Value Models encounter the same problems that were encountered in Our Surprising History and Surprising Cosmology: findings about the distant past or inaccessible regions of space could affect what it would be best to do or permissible to do in implausible ways.

Second, Diminishing Value Models can avoid the consequence that it would be trivial whether or not we extend civilization's future in The Very Last Colony. Even if the final rate of diminished value is very small, the pre-diminished value of an additional good period is very great. So it will never turn out to be inconsequential whether or not human civilization flourishes for another billion years, provided the final rate of diminished value is non-zero.

Third, for related reasons, Diminishing Value Models avoid the consequence that it would be irrational to colonize in The Very Last Colony. They avoid the original problem because, according to these models, there is always some potential upside to colonizing (and thereby producing a good period), even if it is small.

Fourth, Diminishing Value Models avoid Capped Models' implausible implications about The Even Greater Future. For Diminishing Value Models, there is no upper limit to the value of additional good periods of history. Because of this, no matter how long our descendants flourish, it can always be extremely consequential whether our descendants flourish for much longer.

However, Diminishing Value Models do face a problem explaining plausible beliefs about some variations on The Very Last Colony. If the final rate of diminishing value is low, the potential upside of colonization in The Very Last Colony is very small. Yet, it is implausible to "diminish" the importance of the potential downsides. We cannot plausibly claim that since there have been a lot of good periods, it would matter much less if one period were filled with great suffering than it otherwise would have mattered. Because of this, Diminishing Value Models can lead to unreasonable levels of risk aversion. This happens because we are diminishing the importance of the potential upside (perhaps very significantly), without diminishing the importance of the potential downsides. Therefore, if we accept a very low final rate of diminishing value, which we must do if the position is supposed to combat the case for decreasing existential risk, we might end up saying that we should not colonize even if the prospects were quite good. Or, more troublingly, our risk aversion might tell us that some prospects for future generations are better than others, though they seem intuitively worse. This is the period analogue of The Risk-Averse Mother, discussed in chapter 4.

To see how these problems could arise, suppose that the final rate of diminished value were 95% and we were choosing between different prospects for the next period, and the following choices were available:

| | 90% | 10% |
|---|---|---|
| Option 1 | Great period (quality level 100) | Bad period (quality level -100) |
| Option 2 | Neutral period (but people exist) | Neutral period (but people exist) |

If we choose Option 1, there will be a 90% chance that things go very well during the next period, and a 10% chance that things go badly. If we choose Option 2, we will get a "neutral" period for sure. If we discount the upside by 95% and don't discount the downside, it will turn out that Option 2 will be regarded as preferable to Option 1. But this is hard to believe. Some people may reply that we should adopt something like the Tricky Expectation model, discussed in Section 5.1.4.1, in order to explain why Option 1 is better. For reasons analogous to the ones discussed in that section, we should reject this approach.

There is potential for a slightly different reply. Someone might claim that we have reasons of *autonomy*, but not goodness-related reasons, in favor of choosing the Option 1, and that is why it is preferable. Since the people in that period would rationally prefer Option 1, we are required to

honor their wishes, should we wish to ensure their existence.

There are two versions of this reply. On one version, goodness-related reasons *conflict* with reasons of autonomy, but autonomy wins out. On another version, Option 1 is better *in virtue of* the reasons of autonomy in favor of choosing it. Both versions rest on a mistake. They fail because of the Non-Identity Problem. We can set the case up so that if we choose Option 1, different people will exist than if we choose Option 2. Whatever happens, these people cannot truly claim that it would have been in their interest to choose the other option, or that they would have rationally chosen the other option for their own sakes. It still seems that Option 1 is preferable to Option 2, but this cannot be explained via considerations of autonomy.

### 5.2.4 Lost benefits, new challenges

Finally, I can hint at some costs of adopting a Diminishing Value Model, relative to a Capped Model. These issues will be discussed more completely in the next section, and the next chapter.

One challenge, discussed in the next chapter, is that Diminishing Value Models push us toward a certain kind of recklessness. On these models, there is no upper limit to the potential value of adding on good periods, and therefore no upper limit to how good things could be. If there is no upper limit to how good things could be and we are expected value maximizers, then, no matter how good things are, we will be willing to risk everything in order to get a very small chance at something which would be vastly better. And that is hard to accept. On a Capped Model, we can avoid this challenge because we can claim that there is an upper limit to how good things could be. And we do not have to give up on maximizing expected value.

Another challenge is that Capped Models protect us from certain Repugnant-Conclusion-like problems, whereas Period Independence and Diminishing Value Models do not. On Period Independence and Diminishing Value Models, there is great pressure to say that one can make up for losses of quality by just adding enough additional good periods. This is not a logical consequence of these models, but it is very hard to avoid, and it is very hard to believe. I discuss this point more in the next section.

## 5.3 The Repugnant Conclusion and similar problems

### 5.3.1 The problem


Figure 5.2: The Repugnant Conclusion and similar problems

Rather than interpreting the diagram in the usual way, understand height to indicate the quality of a period, and width the number of periods of that kind. So $A$ involves relatively few excellent periods, $B$ involves more periods, though they are not as good, etc. And $Z$ involves very many periods of very low quality.[9]

When they see this case, some people may claim that we must appeal to a Capped Model (and not a Period Independence or Diminishing Value Model) to avoid the claim that $Z$ is better than $A$. In this section, I will argue that these other models are not at a significant disadvantage when trying to negotiate this paradox.

What's paradoxical is that the following things are inconsistent, but each is very plausible:

(5.12) If we have twice as many periods, and things are only slightly less good in each, then things are better. So $B$-worlds are better than $A$-worlds, $C$-worlds are better than $B$-worlds, etc.

(5.13) If things are sufficiently good in each period and we have enough periods, then that is better than any number of periods that are barely worth having. So an $A$-world is better than a $Z$-world.

(5.14) It is possible to transform an $A$-world into a $Z$-world in in a finite number of steps, by doubling the number of periods at each step and making things slightly worse in each period

[9] If we assume that periods have equal numbers of people and well-being is equally distributed across periods in each case, and that the comparative goodness of periods of this kind is proportional to the amount of well-being enjoyed by the people in those periods, this paradox is a version of the intertemporal Mere Addition Paradox.

at each step. We can understand "make things slightly worse" in terms of whatever we think makes things go well during a period.

(5.15) The "better than" relation is transitive.[10]

We can therefore see that it is not just the three models I have proposed that will face these problems—every kind of view will face one of them. Why? Each of the assumptions above is very plausible, but they can't all be true.[11]

### 5.3.2 Potential resolutions

According to the cross-period Capped Model, we should reject (5.12). This is because, after we have had enough periods, it will eventually become (essentially) pointless to ensure that the next period goes well. When this happens, we will not be willing to tolerate a loss in well-being to ensure that the next period goes well. This consequence is, of course, hard to accept. This is relevantly like the case of Voluntary Extinction, discussed in the previous chapter, where people choose to become sterilized in a way that makes their lives go slightly better, at the cost of the end of civilization. And, according to the cross-period Capped Model, this wouldn't be all that bad of an idea, provided there had already been a lot of happy people.

We cannot be too quick to judge Capped Models, since everyone will be drinking some poison here. I have been assuming the transitivity of the "better than" relation throughout, and I will not consider rejecting it here. Therefore, I will only consider rejecting the other assumptions. For a detailed discussion of the costs and benefits of modifying this assumption, see Temkin (2012).

Diminishing Value Models and Period Independence Models, then, have the following options available to them:

They might deny that we can get from $A$ to $Z$ by a finite number of finite expansions in duration and small decreases in quality. There are two sub-options here:

(5.16) Claim that it is impossible to go from $A$ to $Z$ in a finite number of "naturally small" steps.

On this view, if one tried to produce $Z$ from $A$ no finite number of small transformations on

[10] This version of the paradox is constructed using a Spectrum Argument, as in Temkin (2012). For further discussion of Spectrum Arguments, see chapter 7.

[11] As stated, it shouldn't be too clear that Diminishing Value Models fall into this trap. Defenders of these models may claim that we would need the duration to be increased by more before we'd be willing to sacrifice some quality of periods.

This is not an essential problem. We can change the case so that we simply ask the defender of Diminishing Value Models, "By what finite multiple do we need to increase duration before you are willing to sacrifice a tiny percent of quality?" At each stage, we'll give him what he wants. And, sure enough, after a finite number of such transitions, he'll end up moving from an $A$-world to a $Z$-world.

spacetime (extending it, moving particles around, adding particles, etc.), would get you from $A$ to $Z$.

(5.17) Claim that though it is possible to get from $A$ to $Z$ in a finite number of "naturally small" steps, some of these small steps involve extremely significant normative changes.

In unpublished work, Derek Parfit defends (5.17).

Alternatively, we can claim that, sometimes, adding a good period makes things worse, provided it isn't good enough. (On this view, beyond a certain point, doubling and small decreases only make things worse.) These are the analogues of "critical-level" solutions to the Repugnant Conclusion.

Finally, we can accept that $Z$ is better than $A$.

There isn't really any new ground here, just some reminders and colorful commentary. What's important is that almost all of the standard solutions to this kind of paradox, including Parfit's, are open to the defenders of Period Independence and Diminishing Value Models. The only thing they can't say is that, beyond a certain point, more periods don't matter.

As I said in the beginning of this section, it is far from clear which of these solutions takes the least damage, so it is hard to claim that we *must* adopt a Capped Model in response to this paradox, or that we *must* avoid Capped Models. It seems that, at this point, it would be most appropriate to spread our credence out over the space of available solutions, rather than concentrating it on any single point.

## 5.4 Conclusion

We know that we aren't going to find a view that fits with all of our intuitive judgments—even all of the intuitive judgments to which we are deeply committed. It is not enough to look at these models and claim that, since they fail to accommodate our intuitive judgments, they all fail and we should look for some other, entirely different model. Instead, we should summarize our results and attempt to ask which model takes the least damage, and what the most plausible models say about the importance of decreasing existential risk. As in the last chapter, I have produced a table to aid us in this task:

| | The Last Colony | The Very Last Colony | The Even Greater Future | Irrational to colonize the Very Last Colony | The Last Colony vs. The Very Last Colony | The Last Colony vs. The Extra Colony | The Extra Colony vs. The Delayed Colony | Our Surprising History | Surprising Cosmology |
|---|---|---|---|---|---|---|---|---|---|
| Period Independence, spatial cap | | | | | X | | X | | X |
| Period Independence, no spatial cap | | | | | X | X | | | |
| (Early Limits) | | | | | | | | | |
| Capped Model | X | X | X | X | X | X | | X | X |
| DVM 50% final rate, no upper bound | | | | | | X | | X | X |
| DVM 5% final rate, no upper bound | | | | X* | | X | | X | X |
| DVM final rate close to 0, no upper bound | | | | X | | X | | X | X |
| (Late limits) | | | | | | | | | |
| Capped Model | | X | X | X | | | X | X | X |
| DVM 50% final rate, no upper bound | | | | | | | X | X | X |
| DVM 5% final rate, no upper bound | | | | X* | | | X | X | X |
| DVM final rate close to 0, no upper bound | | | | X | | | X | X | X |

X: the view has problems with this case/principle

X*: the view has problems in a version of this case

*The table does not consider the case where we have already reached the limits of diminished value.

Table 5.1: Summary of findings regarding Diminishing Value Models, Capped Models, and Period Independence

Many of these views, especially the ones that took the least damage, do not undermine the case for the overwhelming importance of shaping the far future. By my lights, some version of Period Independence takes the least damage, and this type of view favors the case for the overwhelming importance of shaping the far future. Capped Models and Diminishing Value Models only undermine the case for the overwhelming importance of shaping the far future under special conditions, and the most plausible versions of these views do the least to undermine the case for the overwhelming

importance of shaping the far future. Moreover, these views have their least plausible implications in contexts that involve extinction, such as The Very Last Colony and The Even Greater Future. I conclude that Diminishing Value Models and Capped Models do not present compelling objections to the case for the overwhelming importance of shaping the far future.

# Chapter 6

# A Paradox for Tiny Probabilities of Enormous Values

## Introduction

In this chapter I argue that no moral theory can plausibly deal with tiny probabilities of enormous values. Why believe this? There are two ways to rank prospects under uncertainty, which I call *timid* and *reckless*, there is considerable pressure to choose between them, and each alternative is extremely unappealing. To understand what I mean by *timid* and *reckless*, consider the following two principles:

> *Non-timidity Principle*: If two prospects $A$ and $B$ are similar, except prospect $A$ has a very slightly larger probability of yielding a bad outcome than prospect $B$, but otherwise yields an outcome that is better to a sufficiently great extent, then prospect $A$ is better than prospect $B$.
>
> *Non-recklessness Principle*: If one has the option of getting a sufficiently good outcome $O_1$ with a sufficiently high probability $p_1$, there is some very small probability $p_2 > 0$, such that getting $O_1$ with probability $p_1$ (and getting nothing otherwise) is better than getting an any good outcome $O_2$ (no matter how much better) with probability $p_2$ (and getting nothing otherwise).[1]

If we don't reject one of these principles, then we must violate:

[1] The quantifier ordering here is: $\exists O_1 \exists p_1 \forall p_2 \exists O_2$ such that $(O_1, p_1) > (O_2, p_2)$.

> *Transitivity*: If prospect $A$ is better than prospect $B$, and prospect $B$ is better than prospect $C$, then prospect $A$ is better than prospect $C$.

I explain why this is true in the first section.

In the second section, I explain some relationships between expected utility theory, timidity, recklessness, and moral theories, and some ways in which we can use timidity and fanaticism to evaluate the plausibility of different moral theories. Why does it matter? Unless we're prepared to abandon transitivity, we have to choose between timidity and recklessness, and each position has its own costs and benefits. Given expected utility theory, a moral theory is timid if and only if the theory's utility function is bounded above,[2] and a theory is reckless if and only if its utility function is not bounded above. We can use this information to identify some substantial costs and benefits of adopting different moral theories which have not been adequately appreciated in moral philosophy. This is relevant to the value of shaping the far future because it bears on whether we should accept a Capped Model or Period Independence. Period Independence supports a reckless approach, whereas a Capped Model supports a timid approach. This is true because a Capped Model assumes an upper limit to how good outcomes could be, whereas Period Independence assumes there is no such limit.

In the third section, I argue that, under expected utility theory, timid approaches have very implausible implications in some ordinary cases and some extreme cases. In particular, timidity requires extreme risk aversion when certain extremely good outcomes are possible and extreme risk seeking when certain extremely bad outcomes are possible. Moreover, timid approaches require that events that happened in remote regions of space and time could have relevance to what it would be best to do, even though we cannot affect what happens in those parts of space and time.

In the fourth section, I will argue that ranking outcomes with an unbounded utility function requires making decisions almost entirely on the basis of infinite considerations, in a way that is extremely insensitive to how plausible it is that we could cause any infinitely good or bad outcomes to occur.

In the fifth section, I argue that recklessness poses a severe problem for creating an acceptable theory of decision under moral uncertainty. Philosophers working in the area of moral uncertainty (such as Lockhart (2000), J. Ross (2006a, 2006b), and Sepielli (2010)) are trying to come up with a way of making decisions when one is unsure about morality and which gives some weight to the different moral theories in which the decision-maker has some credence. I argue that recklessness, and the resultant obsession with infinities, is "inherited" under moral uncertainty, so that agents who have

[2] Roughly, a theory's utility function is bounded above if there is some upper limit on how good outcomes can be, according to the theory.

any credence in reckless theories must themselves be reckless, on pain of violating expected utility theory. My argument for this claim makes no assumptions about how intertheoretic comparisons of value are supposed to work.

## 6.1 Why we have to choose

In this section, I argue that we have to choose between timidity, recklessness, and non-transitive rankings of alternatives. A simple example illustrates the point.

> *The Devil at Your Deathbed*: On your deathbed, God hands you a ticket which can be delivered to any of his angels, good for an additional 1 year of happy life, with probability .9999 (otherwise you die now). As you celebrate, the devil appears and asks you, "Would you accept a small risk to get something vastly better? In particular, would you be willing to trade that ticket for one is good for 10 years of happy life, with probability $.9999^2$ (about .9998)?" You accept and the devil hands you a new ticket. Next, the devil asks another question, "Would you accept a small risk to get something vastly better? In particular, would you be willing to trade that ticket for one is good for 100 years of happy life, with probability $.9999^3$ (about .9997)?" After making 10,000 such trades, you get a ticket that would give you $10^{10,000}$ years of happy life if you win, but only works with probability $.9999^{10,000}$, (about one in 20,000). Predictably, you die shortly thereafter.[3]

On one hand, each deal seems like a good one. It seems unreasonably timid to reject a deal which increases your expected length of life by an *enormous* amount, but only increases your risk by an very small amount. On the other hand, it seems unreasonably reckless to take all of the deals—that means trading a really excellent outcome, which you'll enjoy with very high probability, for an *extremely* tiny chance of a vastly better outcome.

Why does avoiding all the reckless choices and all the timid choices lead you in a circle? Consider our options again:

| | 1 | 2 | 3 | 4 | 10,000 |
|---|---|---|---|---|---|
| Payoff | 1 year | 10 years | 100 years | 1000 years | $10^{10,000}$ years |
| Chance | .9999 | .9998 | .9997 | .9996 | $.9999^{10,000} \approx .000045$ |

[3]Temkin (2012) discusses a related spectrum of cases in chapter 8 of his book, *Rethinking the Good*. His spectrum targets the continuity axiom of expected utility theory.

If the Non-Timidity principle holds, then Deal 2 is better than Deal 1, Deal 3 is better than Deal 2, Deal 4 is better than Deal 3,..., Deal 9,999 is better than Deal 9,998, and Deal 10,000 is better than deal 9,999. Then if the "better than" relation is transitive, it follows that all the deals to the left of Deal 10,000 are worse than it. But that violates the Non-Recklessness principle, since that principle says that a very high probability of an excellent outcome can't be worse than an *extremely* small probability of *even better* outcome.

That's why we must choose between timidity, recklessness, and non-transitivity. This chapter will focus on the consequences of taking the timid or reckless horns of the dilemma. I don't discuss the plausibility of rejecting transitivity. For a comprehensive discussion of the reasons for and against accepting transitivity, see Temkin (2012).

The argument does not essentially depend on proportional decreases in probability, or on exponential growth in payoffs. We could have payoffs grow at a cubic rate (1, 8, 27, 64, etc.) and have probabilities drop at a linear rate ($1 - 1 \times 10^{-8}$, $1 - 2 \times 10^{-8}$, $1 - 3 \times 10^{-8}$, etc.). There are an enormous number of possibilities. All that matters is that the payoffs are growing very fast, the probabilities are falling very slowly in comparison, and the probabilities eventually get arbitrarily small (without reaching zero).

Someone might try to embrace the timid approach on the grounds that, after enough time, additional years of happy life have no value (perhaps because of boredom).[4] I don't believe that happy life loses its value after a certain point, but even if it were true, it wouldn't get to the heart of our problem. This problem will arise for other goods as well. If you think that (i) it is better for human civilization to flourish for more time rather than less, and that (ii) if human civilization flourishes for much longer, that would be much better, then you face a variant of the same problem. In general, the problem arises for any good if (i) more of that good is better than less of that good, and (ii) a lot more is a lot better. We can then ask: given a small, fixed, proportional decrease in probability of reward, if you improve a potential payoff enough, is it better to take the additional risk? Always answering yes leads to reckless choices. Ever answering no implies timidity. Doing neither requires ranking the alternatives in a circle.

It's worth noting that we can run a perfectly analogous version of this paradox using *negative* value. In this version, the payoffs are instead years of miserable life, rather than years of happy life. In that version of the paradox, the "timid" person passes up a deal that would make them have a *slightly* larger chance of avoiding misery, but the misery would last *much* longer. In many ways, this is even more implausible than standard timidity. A person who is "reckless," in the negative

[4] For such a view, see Williams (1973).

variant, prefers a very high probability of a very long period of misery to a very low probability of an extremely long period of misery.

Recklessness also bears on the importance of reducing existential risk. One reason it may seem strange to try to reduce existential risk is that trying to avert human extinction seems like such a long shot that it may seem analogous to choosing the devil's 10,000th ticket. And we might hypothesize that whatever is wrong with choosing the devil's 10,000th ticket is the same as whatever is wrong with the view that reducing existential risk is extremely important.

## 6.2 Background and motivation

### 6.2.1 Expected utility theory and bounded/unbounded utility functions

Expected utility theory is the dominant normative model for decision-making under uncertainty. According to expected utility theory, it is rational to rank prospects in accordance with their expected utility. The expected utility of a prospect is computed by (i) identifying all of the outcomes that could follow from choosing that prospect, (ii) assigning a utility to each outcome, (iii) assigning a probability to each outcome, (iv) multiplying the probability of each outcome by its utility, (v) and adding all those terms together to calculate the expected utility of the outcome. The utilities of different outcomes are encoded by a *utility function*, and the probabilities of the different outcomes are encoded by a *probability function*.

At first glance, it may appear that expected utility theory requires agents to act in a reckless way. After all, suppose we start with a ticket for an excellent outcome, $A$, which has a utility of 100.

1. For any small probability $p$, there is an outcome $B$ with a utility greater than $100/p$,
2. Getting $B$ with probability $p$ (and zero utility otherwise) has greater expected utility than getting $A$ for sure.
3. Therefore, getting $B$ with probability $p$ (and zero utility otherwise) is better than getting $A$ for sure.

That argument fails because the first premise may be false. That premise is guaranteed to be true if and only if there is no upper bound to how much utility an outcome can have, but there could be an upper bound. However, the argument does get us somewhere: it shows that all *unbounded* utility functions recommend a reckless approach to decision-making.

What about bounded utility functions? Using a bounded utility function is timid because no matter how high the upper bound is, multiplying that upper bound by a sufficiently small probability results in a very small expected utility. Because of this, for any given prospect with a small probability of getting an enormous reward, the value of that prospect cannot exceed the probability of the reward times the upper bound of the utility function. Perfectly analogous arguments show that timidity in "negative" cases requires a lower bound on the utility function, and recklessness in "negative" cases requires no lower bound on the utility function.

We can illustrate the difference between unbounded and bounded utility functions by considering the following graphs of indifference curves for reckless and timid theories:

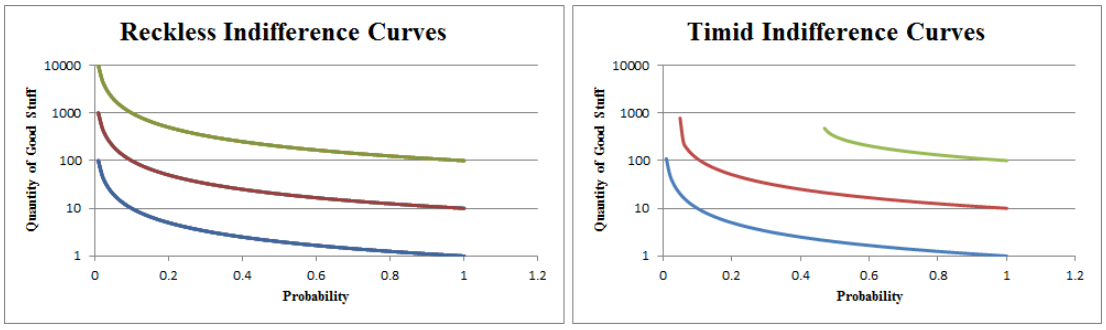


Figure 6.1: Reckless and Timid indifference curves

Here, we have indifference curves over probability and amounts of some good stuff (such as years of happy life, or number of happy people, or duration that human civilization flourishes). Our indifference curves are sets of (probability, quantity of good stuff) pairs that are equally good as prospects. If one prospect pays off with higher probability and pays off better when it actually pays off, it's better. Because of this, the curves represent sets of prospects which are better and better as you you go up and to the right. With the reckless/unbounded indifference curves, for any prospect that yields one outcome with high probability, there is another prospect which yields some much better outcome with much lower probability, such that the two prospects are equally good. In contrast, this is not true with the timid/bounded indifference curve. On these curves, there are some prospects which have a high probability of yielding a very good outcome, and there is no equally good prospect which pays off with a very, very small probability. This is clearest if you look at the upper right curve on the chart depicting timid indifference curves. On that curve, there is no prospect which pays off with probability less than .5 which is as good as the other prospects on that curve.

### 6.2.2 Unboundedness and additive separability

If a moral theory provides a ranking of prospects and the theory is consistent with expected utility theory, then that theory's ranking can be represented by a utility function. Certain moral theories, such as total utilitarianism, must be represented by unbounded utility functions.[5] To see this, note that if we just keep creating additional people with a given level of well-being, the total well-being in the universe will increase by a constant amount each time. If we create enough of these people, the total well-being in the universe can exceed any given finite amount. Something analogous is true of total prioritarianism, though it is less obvious. In that case, (i) making any individual's priority-adjusted well-being greater by a certain amount improves the world to some extent, and (ii) the extent by which the world is improved is independent of the priority-adjusted well-being of other individuals. In economics jargon, total utilitarianism and total prioritarianism are additively separable. A theory is additively separable whenever there are some number of dimensions along which two outcomes or prospects can compare and there are some measures of goodness along those dimensions such that (i) improving the outcomes or prospects along one of these dimensions by a certain amount makes the prospect better to some extent, and (ii) the extent by which the prospect is improved is independent of how good it started out along any of these dimensions. If a theory is additively separable and its ranking of prospects satisfies expected utility theory, then the theory's ranking of outcomes can be represented by a utility function that is a weighted sum of the goodness of prospects along different dimensions.

If there are infinitely many such dimensions, or any given dimension of goodness can take values without upper or lower limits, then the ranking must be represented by an unbounded utility function. Total utilitarianism can have both properties. The dimensions of goodness are the levels of well-being of different people. There are infinitely many potential people, so there are infinitely many dimensions of goodness. And, at least on classical utilitarianism, there is no upper bound to any dimension of goodness, since there is no upper limit to how much well-being any one person can enjoy. Total prioritarianism is also additively separable and unbounded, but in its case the dimensions of goodness are the priority-adjusted levels of well-being of different individuals. For analogous reasons, many additively separable theories of value must be understood using unbounded utility functions.

[5] This is a slight oversimplification, but it is sufficient for our purposes. I am assuming these theories rank prospects in terms of their expected goodness or expected "choiceworthiness." As I use the term, the choiceworthiness of an outcome is how much reason there is to choose the outcome. There is actually a technical wrinkle here which would be of interest to a very small proportion of people who would read this chapter. For an excellent detailed discussion of this technical wrinkle, see Broome (1991, pp. 142-148).

Other, rather different, theories might also be additively separable. For instance, we might accept an additive separability claim about the value of civilizations on different planets, or over different periods of time. This would be to claim that the value contributed to the history of the world by what happens on one planet (or one period of time) is independent of the value contributed to the history of the world by what happens on other such planets (or other such periods of time). If we accept these additive separability claims, we'll also end up with an unbounded, and therefore reckless, ranking of outcomes, since there could be infinitely many such planets or periods of time.

Some rankings of prospects are not additively separable. For example, consider average utilitarianism. On this view, the most salient dimensions along which prospects compare are the well-being levels of the individuals in the prospects compared. This ranking is not additively separable because even though making things better with respect to one of these dimensions always improves a prospect, the extent to which the prospect is improved depends on the state of other dimensions (in particular, how many other people exist).[6]

Even if our all-things-considered ranking of prospects is not additively separable, the ranking may be a function of how good different prospects are with respect to certain ideals, such as well-being, perfectionism, equality, absolute justice, comparative justice, etc.[7] How good a prospect is with respect to a given ideal may be additively separable. Therefore, thinking about additive separability may inform us about the structure of theories that are not themselves additively separable, provided that some of the ideals that matter according to that theory can be understood in additively separable terms.

### 6.2.3 Which normative theories are timid/reckless?

We've established that some important views are additively separable and therefore must be represented by unbounded utility functions. Now that we know this, a new way of assessing these views is to examine the consequences of bounded or unbounded utility functions, and assess whether we find these consequences plausible. Views in the unbounded category include: total views in population ethics, and variants on those views (such as critical-level views); additively separable theories of self-interest, such as classical hedonism; some theories of diminishing marginal value of population (such as average utilitarianism, Hurka's "variable value view" (1983), Ng's Theory X' (1989), and Sider's GV principle (1991)), when combined with unbounded utility functions for individual

[6]The case of average utilitarianism illustrates that although additive separability can imply unboundedness under certain conditions, failure of additive separability does not imply boundedness. This follows because there is no upper bound to how good an outcome can be according to average utilitarianism, provided there is no upper bound to how good an outcome can be for an individual.

[7]I borrow this list of ideals from Temkin (2012, especially ch. 10-12).

well-being; and some Person-Affecting Views, when combined with unbounded utility functions for individual well-being. Views in the bounded category include some bounded theories of self-interest, where Bernard Williams's (1973) view on immortality may be such an example; some theories of diminishing marginal value of population and some Person-Affecting Views, when combined with bounded utility functions for individual well-being.

### 6.2.4 Relevance for the value of shaping the far future

Period Independence is a type of additive separability assumption, and, as we saw in the previous chapter, it places no upper limit on how good outcomes could be. Because of this, when combined with expected utility theory, Period Independence implies recklessness. On the other hand, Capped Models place an upper limit on how good outcomes can be, and therefore, given expected utility theory, they imply timidity.

## 6.3 The Price of timidity

We saw in the last section that if an agent wants to avoid recklessness and satisfy expected utility theory, his ranking of prospects can only be represented by a bounded utility function. In this section I'll discuss the costs of this approach.

### 6.3.1 Violations of Period Independence

One problem is that bounded utility forces us to violate Period Independence, which leads to counter-intuitive implications in cases like Our Surprising History and Surprising Cosmology, as we discussed in the last chapter.

### 6.3.2 Extreme risk aversion in very positive outcomes

Another problem with employing a bounded utility function is that it gives deeply implausible results when thinking about prospects that mix high-value and middling-value outcomes. For example, suppose we ask how many years of life it takes before you get close to the upper bound of your utility function. You must offer some finite number, or else you will not have a bounded utility function. Let's suppose you answer, "10 trillion years." (The number here will be irrelevant.) Now consider two gambles:

| | Heads | Tails |
|---|---|---|
| Deal 1 | 10 trillion years of happy life | 1 hour of misery |
| Deal 2 | $10^{80}$ years of happy life | 2 hours of misery |

Since 10 trillion years of happy life is very close to the upper bound of your utility function, the additional 67 orders of magnitude of years of happy life justifies taking hardly any additional risk at all. However, there is a non-trivial difference between two hours of misery and one hour of misery. For these reasons, the sort of utility function that we have described forces you to prefer Deal 1 to Deal 2.

### 6.3.3 Extreme risk seeking in very negative outcomes

A utility function that is bounded from below faces even more troubling problems. As we saw in Section 1, there is an analogous decision to make between "negative" timidity and "negative" recklessness. An argument parallel to the one given in Section 3.2 forces a similar conclusion about the following prospects:

| | Heads | Tails |
|---|---|---|
| Deal 3 | 10 trillion years of misery | 1 hour of happy life |
| Deal 4 | $10^{80}$ years of misery | 2 hours of happy life |

If 10 trillion years is close to the lower bound of your utility function, you must prefer Deal 4 to Deal 3. Again, this is very hard to believe. The consequences are even more troubling when we consider interpersonal cases that involve suffering. Analogous assumptions lead to the conclusion that timid agents must prefer Deal 6 to Deal 5 in an interpersonal case of the following kind:

| | Heads | Tails |
|---|---|---|
| Deal 5 | 10 trillion people suffer for a year | 1 person has an enjoyable day |
| Deal 6 | $10^{80}$ people suffer for a year | 1 person has two enjoyable days |

Other than expected utility theory and timidity, the assumptions that lead to this are that: (i) more people suffering is worse than fewer people suffering, (ii) 10 trillion people suffering is close to the lower bound of the utility function, (iii) it is better for one person to have more enjoyable days of life, other things being equal, and (iv) one enjoyable day is not close to the upper bound of the utility function. Analogous results follow if a different lower bound is chosen. Given the plausibility of the background assumptions, it seems that we can blame timidity for these strange conclusions about gambles.

### 6.3.4 Itemized billing for the timid

To summarize, timid agents face the following difficulties.

1. They pass up at least one seemingly great low-risk deal in The Devil at Your Deathbed.
2. If they rank prospects using expected utility theory...
   (a) Their ranking of prospects is sensitive to seemingly irrelevant differences in the far removes of time and space, in a way that violates plausible additive separability requirements.
   (b) They rank prospects in an extremely risk averse way when certain extremely good outcomes are possible.
   (c) They rank prospects in an extremely risk seeking way when certain extremely bad outcomes are possible.

Though these problems seem pretty bad, we should compare the costs of being reckless. We already know that every option is going to have some pretty unpalatable implications.

## 6.4 Recklessness and fanaticism

We've already identified one problem with recklessness: willingness to take extreme risks, even in cases where the default outcome is very good. In defense of the reckless, we can say a few things. It is hard to comprehend goods that are getting arbitrarily large, so perhaps intuition is not to be trusted in this situation.[8] So far, we've seen no argument that the reckless will spend their time obsessing over far-fetched scenarios involving unreasonably large finite goods unless the probabilities work out right, which they may not in practice. Perhaps in extreme situations where the probabilities do work out, reckless behavior would be reasonable. Looking at the current score, recklessness seems less costly than timidity. However, recklessness is not consistent with certain received wisdom about the possibility of infinitely valuable outcomes. Many philosophers seem to be attracted to the following package of views about infinities:

1. In at least some cases, an infinite amount of something good is better than any finite amount. For instance, being happy forever is better than being happy for any finite duration, and human civilization flourishing forever is better than human civilization flourishing for a finite duration.

[8] See Baron and Greene (1996) for enlightening examples about how pre-theoretic intuition fares very badly with respect to evaluating alternatives involving large numbers of people.

2. There are severe theoretical problems with assigning value to infinite outcomes and infinite prospects. Someday, new mathematics or philosophy might solve these, but at the moment, it is impossible to say much of general importance about how to handle small probabilities of infinite value.

3. In practice, infinite considerations have little decision-theoretic weight. If an infinite consideration–such as the possibility of living in heaven forever–is sufficiently far-fetched and there is no empirical evidence in its favor, it is rational to ignore that possibility. All infinite considerations are like this, so that infinite considerations have no practical consequences.

In contrast, some philosophers, such as Pascal, would claim that infinite considerations should dominate finite considerations, at least in theory. More precisely, these people claim:

> Fanaticism: Any non-zero probability of an infinitely good outcome, no matter how small, is better than any probability of a finitely good outcome.

In the next section, I illustrate the distinction between mere recklessness and fanaticism; argue that given certain minimal assumptions, reckless agents must be fanatical, on pain of inconsistency; and argue that fanaticism could have practical consequences in some real cases.

### 6.4.1 Recklessness vs. fanaticism

How is recklessness different from fanaticism? I'll clarify that in a second. First consider the following story:

> *God's Shop*: While visiting God's shop of goods that are guaranteed not to be fake, the descriptions of two items catch your eye:
>
> "Extend your life by 1 million years! Just $20! Disclaimer: this product has a 1% failure rate."
>
> "Get a chance to extend your life by 1,000,000____ years! Only $20! Disclaimer: only works in 50% of cases."
>
> The description of the second item is faded, and you can't tell how many zeros there are. There could be a lot of them. Unfortunately, you only have $20 in your pocket and God doesn't take loans or credit cards. What do you do?

A timid person might choose the first item or the second, depending on how high the person's upper limit on the value of additional years of life would be. If 1 million years is close to the upper limit,

he wouldn't bother asking the shopkeeper how many zeros were in the second item's description, since the second option couldn't be more than twice as good, and therefore he knows that the first prospect is better than the second. A reckless person would ask the shopkeeper how many zeros were on the second item's description, wait patiently for the shopkeeper to check the books, and then might go for the second item, depending on how many zeros there were. A reckless person is willing to take an extreme risk, but only if the potential payoff is sufficiently large.

Consider a variant on God's Shop:

> *God's Shop, Three Options*: Like God's Shop, but there is a third item description:
>
> "Get a chance at living forever! Disclaimer: only works in 1 in 1,000,000,000,____ cases. Get yours for just $20!"
>
> Once again, the description is faded, and you can't tell how many zeros are there.

A timid person would, once again, stick with the first option, and not bother looking at the other two (assuming that 1 million years is close enough to the upper limit of how good the outcome could be). The definition of recklessness does not immediately imply anything about how the third option compares to the second. A fanatic, on the other hand, would go for the third option, *regardless of how many zeros were* on any of these product descriptions, since, for them, any non-zero chance of an infinitely valuable outcome dominates all finite considerations.

That's one major difference between the reckless and the fanatical: though the reckless are willing to take some extreme risks, their willingness to take these risks depends on the probability of the risk paying off; fanatics have no such concern. For any particular longshot that a reckless person is interested in, there's some conceivable evidence you could show him, short of definitively *proving*, with probability 1, that his longshot won't pay off, and if you do it, that will make him give up his interest in the longshot. Not so for a fanatic; unless you definitively prove, with probability 1, that his infinite longshot won't pay off, he'll prefer to take his chances.

Another difference is that fanaticism implies a violation of the continuity axiom of expected utility theory on its face, whereas you have to do a bit of work to see that this is true for a reckless agent (we'll do that work in the next subsection). The continuity axiom says that for any three outcomes $A$, $B$, and $C$, where $A$ is preferred to $B$ which is preferred than $C$, there is some probability $p > 0$ such that getting $A$ with probability $p$ and $C$ with probability $1 - p$ is ranked equally as high as getting $B$ for sure. Fanatical agents violate this rule, since they treat some infinitely good outcomes as better than any chance of a finitely good outcome. For example, let $A$ be going to heaven and being happy forever, let $B$ be the best life any mortal has ever lived, and let $C$ be a normal human

life. For any probability $p > 0$, no matter how small, a fanatical agent prefers ($A$ with probability $p$ and $C$ with probability $1 - p$) to getting $B$ for sure.

### 6.4.2 How recklessness leads to fanaticism

Unfortunately, being reckless leads to fanaticism, given only very weak assumptions. For example, suppose we have an agent who is reckless about the number of happy years of life he has. Suppose that for any number of years of happy life and any high probability of enjoying those years, he'd prefer a very small chance of living for a much longer time. If that agent's preferences satisfy transitivity and the Sure-Thing principle,[9] he'll prefer any chance of infinitely many years of happy life to any finite number of years of happy life. If such an agent gives even the slightest credence to the possibility that doing something would make him live forever, he'll pursue it at any finite cost, no matter how large. In other words, he'll be fanatical.

This can be proven as follows. Let $X$ be an outcome where the reckless agent lives forever. We'll show that for any number, n, of years of life and any probability $p > 0$, the reckless agent must prefer living forever with probability $p$ (and dying immediately otherwise) to living n years for sure. So let $n$ be given. Recall that for any number, $n$, of years of life and any small probability, $p$, there is some much larger finite number, $k$, of years of life such that a reckless agent prefers getting $k$ years with probability $p$ to getting n years for sure. If the agent prefers living forever to living for any finite amount of time, he will prefer living forever with probability $p$ to living for $k$ years with probability $p$ (this follows by application of the Sure-Thing principle), and will therefore prefer living forever with probability $p$ to living for $n$ years for sure (this follows by application of transitivity). Since $n$ was arbitrary, it follows that for any number of years of life $n$ and for any $p > 0$, the agent prefers living forever with probability $p$ to living for $n$ years for sure, which was what we wanted to show.

That last paragraph is actually a mathematical proof, but it isn't a very intuitive one. What's really going on here? Say Outcome $X$ is our infinitely good outcome. When you've got a reckless agent, they're willing to trade any finitely good deal they have for a very tiny probability of something else, as long as it would be *a lot* better if they get it. That means that no matter how low of a probability you name, there's some finite longshot that pays off with that much probability that the agent regards as better than whatever he started with. Outcome $X$ is a lot better than any finitely good outcome. So whenever someone would go for a finite longshot, they should prefer a longshot that gives them Outcome $X$ if it pays off. This is true *for any* small probability, so the agent has

[9] Basically, the Sure-Thing Principle tells you to follow rules like: prefer a gamble that gives you \$5 if you roll an even number to a gamble that gives you \$5 if you roll a six. More generally, it tells you to prefer one prospect to another whenever the first prospect could yield a better outcome and couldn't yield a worse outcome.

to prefer any small probability of Outcome $X$ to whatever he started with. But we didn't make any assumptions about what he started with, except that it was finite, so the agent regards any probability of Outcome $X$ is better than anything finitely good.

We just made an argument for fanaticism along the dimension of duration. But there are other ways fanatical agents might obsess over infinite longshots. If some pleasure were infinitely good over some period of time, that could induce fanaticism, and likewise for a good that was enjoyed by infinitely many people. And if an agent had any finite chance of playing a St. Petersburg-like gamble, that could induce fanaticism as well.[10][11] More generally, if any ranking is additively separable across some dimensions of goodness and one of these dimensions of goodness could be infinitely great, prospects where that dimension is infinitely great will induce fanaticism.

### 6.4.3 Does fanaticism have practical consequences?

The theoretical consequences of fanaticism are pretty bizarre. I don't think many of us are happy to say that what actually makes a decision best is whether it optimizes for infinite considerations, regardless of how little infinite consideration probability is at stake. It's less clear how strange the practical consequences are. In this subsection, I consider three arguments that fanaticism has limited practical consequences.

#### Having zero probability in all possibilities for achieving infinite value is implausible

The most obvious way to avoid any practical consequences of fanaticism is to have probability 0 in any scenario where you achieve infinite good. Call a person with such probabilities *dogmatic*. The trouble with this approach is that it seems irrational to only assign probability zero to all the infinite scenarios that have been contemplated by physicists, futurists, and theists. Having probability zero

[10] In the St. Petersburg gamble, there is a ½ probability of getting \$1, a ¼ probability of getting \$2, a 1/8 probability of getting \$4, etc., so that the gamble returns an infinite expected amount of money. The St. Petersburg gamble doesn't have infinite expected utility because there is an upper bound to how much utility an agent derives from increasing amounts of money. But it is pretty easy to show that for any unbounded utility function it is possible to construct a St. Petersburg-like gamble which must outrank any outcome with finite utility. The trick is just to find outcomes $O_1$, $O_2$,..., where outcome $O_i$ with utility greater than or equal to $2^i$ and probability equal to $2^{-i}$. According to Paul Samuelson's history of the St. Petersburg paradox (1977), this was first observed by the mathematician Karl Menger in 1934.

[11] Arrow (1971) observed that having an unbounded utility function over outcomes was inconsistent with the continuity axiom because of the possibility of Menger's St. Petersburg-like gambles, which were discussed in a footnote above. Such a gamble is ranked above any of those outcomes over which the unbounded utility function is defined, which is inconsistent with the continuity axiom. This result can seem puzzling, since it is natural to think that it only makes sense to rank outcomes in accordance with a utility function if one accepts expected utility theory, and expected utility theory assumes the continuity axiom. The tension disappears if one remembers that it is logically consistent to hold that the assumptions of expected utility theory apply in some cases, but not in others. If we take a set of alternatives where expected utility theory applies and ranks the alternatives with an unbounded utility function, extending that ranking to a larger set of alternatives may violate some of the assumptions of expected utility theory. As I'm making the argument, continuity applies for ranking prospects with only finitely many possible outcomes, all of which have finite value, but continuity sometimes fails when we bring in prospects with infinitely many possible outcomes.

in these claims would require being willing to take a bet that would win you a penny if you were right, but lose you your life savings if you were wrong. Few people who claim certainty about these issues would accept these bets, so I doubt they are truly certain that all infinite scenarios are impossible.

**Not all infinite expectations are equal**

Another argument that fanaticism has no practical consequences goes as follows:

1. If doing $A$ has infinite expected value, then a prospect with any probability of resulting in $A$ has infinite expected value.

2. All alternatives with infinite expected value are equally good in terms of infinite considerations.

3. For almost any two things we might do, each has some probability of producing an infinitely valuable outcome.

4. Therefore, for almost any two things we might do, both are equally good in terms of infinite considerations.

For example, if praying to God has infinite expected value, so does flipping a coin to decide whether to pray to God, or even just living in a country where you might pray to God at some point or other.[12] So giving yourself a non-zero chance of pursuing some infinite consideration gives you infinite expected value as well.

We should reject the assumption that all prospects with infinite expected value are equally good in terms of infinite considerations. On a purely intuitive level, going to heaven for sure is obviously better than going to heaven with probability one in a million. Moreover, expected utility theory provides no good argument for the contrary view. In fact, one axiom of expected utility theory, the Sure-Thing Principle, is inconsistent with the contrary view. The Sure-Thing Principle says that if one prospect might lead to a better outcome than another, and couldn't lead to a worse outcome, it's a better prospect. That implies that going to heaven for sure is better than going to heaven with probability one in a million. Since there is no argument in favor of this assumption from expected utility theory, the assumption is intuitively implausible, and the assumption is inconsistent with a very plausible general principle, we should reject this assumption.

[12] Hajek (2003) discusses this thought more sympathetically than I do.

**Empirical stabilizing assumptions may work in some cases, but won't work in general**

Another way to try to avoid practical consequences is to invoke what Bostrom (2011) calls an empirical stabilizing assumption—some empirical assumption which, if true, "do what is best in terms of finite considerations" and "do what is best in terms of infinite considerations" are broadly equivalent. One can imagine other domains where something analogous holds. For example, it may be that for many children "do what your mother says" and "do what is in your own best interest" are broadly equivalent, given that these children's mothers only tell them to do what is in their best interest.

What kind of empirical stabilizing assumption might someone hold? Someone might argue that "do what best promotes the long-term survival and flourishing of our descendants in the far future" is broadly equivalent with "do what is best in terms of finite considerations" and "do what is best in terms of infinite considerations." It may be best with respect to finite considerations for the reasons I argued in earlier chapters of this dissertation. Perhaps it's the best with respect to infinite considerations because if humans manage to survive into the distant future, they're likely to find and execute any available strategies for achieving infinitely good outcomes. Because future people are likely to be much more numerous and capable than us, it is plausible that they could execute these strategies on a much, much greater scale than we could.

Reflecting on an analogy may make the second claim more plausible. Suppose that in 1500 CE, someone wrote a forward-looking novel that featured a technology from the present day, such as a telephone. And suppose another person read this novel and then set for himself the goal that, in the future, people utilized rapid long-distance communication as effectively as possible. He would know that if making telephones was actually a good idea, future people would be in a much better position to find a way to create telephones and use them effectively. He would know very little about telephones or how they might be discovered, so it would not make sense for him to do something very targeted, such as drafting potential telephone designs. It would make more sense, I believe, for him to help in very broad ways (such as becoming a teacher or fighting political and religious threats to the advance of science), thereby empowering future generations to discover and effectively utilize rapid long-distance communication.

For this story to work, it's important that (i) future generations will have an interest in rapid long-distance communication if it is possible at all, (ii) future people will be vastly better at creating and utilizing rapid long-distance communication if it's possible at all, and (iii) there are currently not very effective ways to create and utilize rapid long-distance communication directly. Analogously,

it seems that (i) future generations will have an interest in producing infinitely good outcomes if it is possible at all, (ii) future people will be vastly better at producing infinitely good outcomes if it's possible at all, and (iii) there are currently not very effective ways to produce infinitely good outcomes directly. For reasons like these, the above empirical stabilizing assumption seems plausible.

Therefore, fanaticism may not change our most important altruistic decisions in a significant way, except perhaps to significantly increase our interest in assessing whether candidate empirical stabilizing assumptions hold. But the assumption is likely to fail in some cases we may encounter in the distant future:

> *Infinite Research vs. Utopia*: Our descendants reach the limits of technological progress and become very convinced (with probability $1 - 10^{-N}$, for some really huge $N$) that achieving an infinite amount of good is impossible. They must decide how some vast amount of resources should be allocated between two projects: creating an extremely good (though only finitely good) utopia, or researching possible methods of achieving an infinitely good outcome.

It is likely that if these people were fanatical, they would spend nearly all of the resources on the infinite research, they would keep becoming more and more certain that their research would bear no fruit, and they would keep doing this as long as they didn't become completely certain that achieving an infinitely good outcome was impossible, which would never happen.

### 6.4.4 But that's an infinite case!

Some defenders of the reckless might respond with some variation of, "Ah yes, but that's an infinite case, and those are paradoxical for everyone." The first thing to say is that infinite cases can be handled in mathematically consistent ways, and it is common to produce physical theories with infinite domains. There is no particular reason to hope that there will be some revolution in the mathematics of infinity which solves all of the infinity-related problems in ethics and decision theory.

The second thing to say is that infinite cases are not necessarily esoteric. As Bostrom (2011) has pointed out, most astrophysicists now believe that the universe is infinite, and this makes it very likely that there is an infinite quantity of valuable stuff. If we want a fully general theory of the value of prospects, we can't just "bracket" the issue of infinity forever.

### 6.4.5 Wrapping up

In this section I've argued for the following claims. Rational agents must be fanatical if they are (i) reckless, (ii) non-dogmatic, (iii) prefer the infinite to the finite, (iv) and have preferences that are transitive and satisfy the Sure-Thing principle. Fanaticism is implausible, and is the major cost of being reckless. It's not clear what the practical consequences of fanaticism would be. Empirical stabilizing assumptions may imply that accepting fanaticism would have only minor effects on our views about which prospects are best in many ordinary situations, though there would probably be some major differences in some cases we are likely to encounter. It is impractical to dismiss infinite considerations by claiming that we should simply "bracket" infinite cases forever.

## 6.5 Fanaticism and decision under moral uncertainty

Recently, some philosophers have discussed the question of how it would be appropriate for individuals to take account of their uncertainty about morality when they decide what to do.[13] Target questions would include the following:

1. If someone has some credence in utilitarianism and some credence in a pluralistic form of egalitarianism, how much weight should that person give to promoting equality in light of his uncertainty?

2. If I have 20% credence that Peter Singer (1972) is right about our obligations to the needy, should I start donating in order to hedge my bets, morally speaking?

In this section, I argue that the possibility of reckless and fanatical theories poses a significant challenge for constructing an adequate theory of decision under moral uncertainty.

Jacob Ross (2006a, 2006b) and Andrew Sepielli (2010) have argued that in conditions of moral uncertainty, we should maximize expected value, relative to our uncertainty over normative theories. As discussed by Ross (2006b), it is natural to object that this kind of strategy gives too much weight to extreme, low probability theories. The worry is that we might be almost sure that such theories are false, but have to act as if they were true because of expected utility considerations. Some of the ideas developed in this chapter suggest that meeting this challenge will have severe costs and that there will be special challenges for avoiding recklessness and fanaticism when developing a theory of moral uncertainty, at least given the background assumption of expected utility theory.

[13] For background, see Lockhart (2000), Ross (2006a), Ross (2006b) and Sepielli (2010).

To see the issue, suppose we rank prospects relative to our moral uncertainty in an additively separable way, where the "dimensions" of goodness are the strengths of the moral reasons in favor of choosing the different prospects, according to the different moral theories we accept. Given expected utility theory, these different dimensions will be representable by utility functions that correspond to each of the moral theories one has credence in. As I pointed out in Section 2, if any of these dimensions is representable by an unbounded utility function, then the all things considered ranking must be representable by an unbounded utility function.[14] And, as I pointed out in the last section, if some ranking is additively separable along multiple dimensions and it is possible for any one dimension to be infinitely good, the overall ranking must be fanatical. The consequence is that if we have any credence in theories that are themselves reckless or fanatical and our overall ranking of prospects is additively separable across the different theories in which we have credence, then our overall ranking of prospects is itself reckless or fanatical.

On Ross and Sepielli's approach, the goodness of prospects relative to our moral uncertainty is additively separable with respect to the different theories in which we have credence, so their approaches must be reckless/fanatical if they allow any credence in reckless/fanatical theories. Despite the drawbacks of recklessness and fanaticism, it is hard to deny that we should have some credence in reckless/fanatical views, given the challenges faced by the alternatives, and given that many theories that many serious people have defended are reckless/fanatical. It is therefore hard to deny that these approaches will lead to reckless/fanatical implications.

One might think, "So much worse for that type of additively separable approach!" The trouble with this reaction is that that type of additively separable approach is strongly supported by expected utility theory. As I show in an appendix, an additively separable approach is mathematically inevitable given the following assumptions:

1. *Expected Utility Assumption (for moral theories)*: The theories in which one has non-zero credence rank prospects using utility functions.

2. *Expected Utility Assumption (for decision under moral uncertainty)*: One's ranking of prospects (relative to one's moral uncertainty) satisfies expected utility theory.

3. *Pareto Assumption*: The ranking of prospects (relative to one's moral uncertainty) prefers prospects that are better according to some theories and worse according to none.

[14] There would be some cases where certain theories would exactly cancel out and this would not be true, but it would not be the norm. For example, if total utilitarianism and the opposite of total utilitarianism were the only unbounded moral theories in which one had credence, and one had equal credence in each of them, then one's overall ranking relative to one's moral uncertainty would be bounded. But, in general, it would be a bizarre miracle if all one's uncertainty in unbounded normative theories canceled out exactly.

The result follows directly given a simple reinterpretation of Harsanyi's Aggregation Theorem (Harsanyi 1955), and I provide a proof in the appendix.

Someone might try to avoid these difficulties by appealing to the problem of intertheoretic comparisons of value. An additively separable approach requires adding up utility functions representing the values of different moral theories. And it may seem that there is no common scale along which, for example, 5 units of total-utilitarian value can be compared with 25 units of average-utilitarian value. After all, with any utility function, we could have just as easily represented its ranking of prospects by multiplying all of its values by any positive number. The beauty of the argument I've presented here is that it makes no appeal to intertheoretic comparisons of value; it does not assume that there is any "privileged" way to compare units of value-according-to-total-utilitarianism and units of value-according-to-average-utilitarianism. This is evident if you look at the proof in the appendix.

Another reply to this argument is that the first assumption is far too strong. Even if expected utility theory is true, it would be a mistake to have 100% credence in it. And if we don't have 100% credence in expected utility theory, then some of the theories we have non-zero credence in *don't* satisfy expected utility theory. And therefore, the first assumption doesn't hold in practice. But many people may have substantial credence in theories that *do* satisfy expected utility theory, and my reinterpretation of Harsanyi's theorem tells us something significant about what it would be best to to relative to our moral uncertainty, conditional on expected utility theory. For people who don't have 100% credence in expected utility theory, this won't immediately determine what it is best to do, but it will probably be an input to what it would be best to do relative to our moral uncertainty all things considered.

One final observation is that fanaticism is a special, open problem for philosophers who wish to develop an adequate theory of decision under moral uncertainty. Followers of expected utility theory can avoid all of the difficulties of fanaticism if they just adopt bounded utility functions, though of course they inherit all the difficulties of timid theories. But in the case of moral uncertainty, one should always have *some* remaining confidence in unbounded approaches, and that makes it much harder to avoid fanaticism.

## 6.6 Conclusion

In summary, in order to follow a rational policy, we must be willing to pass up arbitrarily great gains, even at small risks (be timid), be willing to risk everything at arbitrarily long odds for the

sake of enormous potential gains (be reckless), or rank our prospects in a non-transitive way. Given expected utility theory, timid approaches force us to violate Period Independence and lead to strange implications about cases like Our Surprising History and Surprising Cosmology.

In some ways, recklessness seems less bad. In practice, reckless agents need not fanatically obsess over impossibly large finite prospects, since the probabilities might not align properly. But if agents do not dogmatically deny the possibility of infinite good, and they prefer infinite good to finite good, they must be fanatical, pursuing any chance of the infinite at any finite expense.

Some may see this as another argument for ranking prospects non-transitively. I find this approach unsatisfying, but I think that increasing one's confidence in this position is a fair reaction to the arguments I have presented, given that new challenges have been presented for other approaches, but not for this one.

Since ranking systems which use unbounded utility functions lead to recklessness and fanaticism, and ranking systems with bounded utility functions lead to timidity, we now know something about the costs and benefits of the different approaches, and these costs and benefits could be included when we weigh up the plausibility of different moral theories. As we saw in Section 2.3, this has relevance for many different positions in normative theory, especially in population ethics. Reckless theories, such as total views in population ethics, are committed to "obsessing over infinities." That's a surprising fact about theories like total utilitarianism; it's not commonly thought of as a major consideration for assessing whether the theory is acceptable, but it probably should be. And timid theories, such as some theories of diminishing marginal value of population, have all the bad consequences that we've said timid theories have. We should keep such facts in mind when evaluating the comparative plausibility of theories from each of these classes.

Recklessness and fanaticism are special challenges for developing an adequate theory of decision under moral uncertainty. In developing such a theory, we face the uncomfortable choice of (i) abandoning expected utility theory, (ii) claiming to have literally zero credence in any reckless/fanatical theory, (iii) preferring ranking systems that are worse according to some theories we have credence in and better according to none, or (iv) accepting a reckless/fanatical ranking of prospects under moral uncertainty.

All options are deeply unpalatable, so we are left with a paradox. In the next chapter, I offer some suggestions about how it would be best to proceed from a practical perspective, in light of this paradox.

# Appendix: Proof of additive separability claim

In this appendix I precisely formulate and prove my claim that, under certain compelling conditions, systems of decision-making under moral uncertainty will be additively separable across the utility functions corresponding to the different theories in which we have some credence.

Theorem: If the following assumptions hold,

1. Expected Utility Assumption (for individual rankings): There are $n$ different rankings $>_1, \ldots, >_n$ over prospects, and a subset $X$ of these prospects such that, for each $i$, there is a utility function $U_i$ such that for any prospects $A$ and $B$ in $X$, $U_i(A) > U_i(B)$ if and only if $A >_i B$.

2. Expected Utility Assumption (for master ranking): There is a ranking of these prospects, $>$, and a utility function $U$ such that for any prospects $A$ and $B$ in $X$, $A > B$ if and only if $U(A) > U(B)$.

3. Pareto Assumption: For any prospects $A$ and $B$ in $X$, if for every $i$, $A \geq_i B$, then $A \geq B$. If the former holds and there also exists a $j$ such that $A >_j B$, then $A > B$.

Then there are some constants $a_i > 0$ such that for any $A$ in $X$, $U(A) = a_1U_1(A) + \ldots + a_nU_n(A)$.

Proof: To prove the theorem, we'll proceed in a fashion analogous to Harsanyi (1955, pp. 313-314). We'll first prove the following Lemma.

Lemma: For any prospects $A$ and $B$ in $X$, and for any $k$, if for all $i$, $U_i(A) = kU_i(B)$, then $U(A) = kU(B)$.

Proof of Lemma: First, consider the case where $0 \leq k \leq 1$. Let $Z$ be an outcome such that $U_i(Z) = 0$ for all $i$, and $U(Z) = 0$. (Such a prospect is guaranteed to exist because the zero on a utility function is arbitrary.) Let $Q$ be a gamble which returns $B$ with probability $k$ and returns $Z$ otherwise. Then $U(Q) = kU(B)$. And we also know that for each $i$, $U_i(Q) = kU_i(B) = k \times 1/k \times U_i(A) = U_i(A)$. Therefore, for each $i$, $U_i(Q) = U_i(A)$. Therefore, from the Pareto Assumption, we have that $U(Q) = U(A)$. Therefore, we have that $U(Q) = kU(B)$, and that $U(Q) = U(A)$, which implies that $U(A) = kU(B)$.

Second, consider the case where $k < 0$. Let's consider a prospect $R$ that yields $A$ with probability $1/(1-k)$ and $B$ otherwise. For each $i$, $U_i(R) = 1/(1-k) \times U_i(A) + (-k)/(1-k) \times U_i(B) = 1/(1-k) \times kU_i(B) + (-k)/(1-k) \times U_i - (B) = 0$. Since $U_i(R) = 0$ for all $R$, $U(R) = U(Z)$, by the Pareto Assumption. Hence, $U(R) = 1/(1-k) \times U(A) + (-k)/(1-k) \times U(B) = 0$, which implies that $U(A) = kU(B)$.

Third, consider the case where $k > 0$. Let's consider a prospect $S$ that yields $A$ with probability $1/k$ and $Z$ otherwise. For each $i$, $U_i(S) = 1/k \times U_i(A) + (1 - 1/k) \times U_i(Z) = 1/k \times U_i(A)$, which implies that $U_i(A) = kU_i(S)$. Since $kU_i(B) = U_i(A)$, it follows that $U_i(S) = U_i(B)$. Therefore, by the Pareto Assumption, $U(S) = U(B)$. And $U(S) = 1/k \times U(A) + (1 - 1/k) \times U(Z) = 1/k \times U(A)$. Thus, $kU(S) = U(A)$. Together with $U(S) = U(B)$, that implies that $U(A) = kU(B)$. Since we have covered all possible values of $k$, that completes the proof.

Proof of Theorem: Let $Z_i$ be a prospect such that $U_i(Z_i) = 1$, and for any $j$ other than $i$, $U_i(Z_i) = 0$. Let $U(Z_i) = a_i$. Note that all the $a_i$ must be greater than zero, since $U(Z) = 0$ and $Z_i > Z$ by the Pareto Assumption.

Let a prospect $A$ be given. Consider prospects $S_1, \ldots, S_n$ such that, for each $i$, $S_i$ is a prospect where $U_i(S_i) = U_i(A)$, and for any $j$ other than $i$, $U_j(S_i) = 0$. First observe that for all $j$, $U_j(S_i) = U_j(Z_i)U_i(S_i)$. Therefore, by the Lemma we just proved, $U(S_i) = U(Z_i)U_i(S_i)$.

Let $T$ be a prospect that yields one of the prospects $S_1, \ldots, S_n$, where each outcome has probability $1/n$. Let $a_i = U(Z_i)$. Now, for each $i$, $U_i(T) = 1/n \times U_i(A) + (1 - 1/n) \times 0 = 1/n \times U_i(A)$. Therefore, by the Lemma, $U(A) * 1/n = U(T)$. And $U(T) = 1/n * (U(S_1) + \ldots + U(S_n))$. Hence, $U(A) = U(S_1) + \ldots + U(S_n)$. Above, we showed that $U(S_i) = U(Z_i)U_i(S_i)$. Therefore, $U(A) = U(Z_1)U_1(S_1) + \ldots + U(Z_n)U_n(S_n)$. Therefore, there exist constants $a_i > 0$ such that $U(A) = a_1U_1(A) + \ldots + a_nU_n(A)$, which completes the proof.

Some readers may recognize this as Harsanyi's Aggregation Theorem, or Broome's (1991) "interpersonal addition theorem." Mathematically, it's the same result. But it can be given a novel interpretation if we consider the $>_i$ as rankings of prospects under different moral theories in which we have some credence, and $>$ as a ranking of all prospects, relative to one's moral uncertainty.

Under the new interpretation, the first assumption says that all the different theories we have credence in rank the prospects in $X$ in a way that satisfies expected utility theory. The second assumption says that our overall ranking of prospects relative to our moral uncertainty also ranks the prospects in $X$ in a way that satisfies expected utility theory. And the third assumption says that (i) if $A$ is ranked at least as high as $B$ according to all the theories we have credence in, then $A$ is ranked at least as high as $B$, relative to our moral uncertainty; and (ii) if, in addition, $A$ is ranked higher than $B$ according to at least one theory we have credence in, then $A$ is ranked above $B$, relative to our moral uncertainty.

The conclusion says that our utility function $U$, which ranks prospects relative to our moral uncertainty, must be a weighted sum of the utility functions associated with all the theories in which we have some credence. In other words, $U$ is additively separable across the utility functions

corresponding to the different theories in which we have some credence, which is what I wanted to show.

# Chapter 7

# Infinite Value, Long Shots, and the Far Future

## Introduction

In the previous chapters, I have defended a package of views which suggest that shaping the far future is overwhelmingly important. However, that package of views implies *fanaticism*—the view that any probability of any infinitely good outcome is better than any probability of any finitely good outcome. As I showed in the previous chapter, fanaticism has some very implausible implications, but the alternatives have implausible implications as well. This should make us feel both more confused, and more skeptical of the package of views that I defended in earlier chapters.

In this chapter, I try to make us less confused about this, but I start by talking about various aspects of the problem that make it more confusing. In the first section, I highlight some difficulties for saying what would be best with respect to infinite considerations, and explain how what is best with respect to infinite considerations may depend on whether our universe is likely to have an infinite amount of valuable stuff already, and whether we make distinctions between different levels of infinite value. In the second section, I examine how a timid approach to assessing the value of prospects bears on the value of shaping the far future. The right conclusion about this issue depends on many complicated issues, such as how value is aggregated across periods of history, whether there is an infinite amount of valuable stuff in the universe, whether we include events far outside of our causal control when aggregating value across space and time, how we make trade-offs between different types of infinities, and what the upper limit is for how good outcomes can be (if

there is such a limit). In the third section, I consider relying on the first approach where it has plausible consequences, and relying on the second approach where it has plausible consequences. Using these different approaches in different contexts will result in ranking options non-transitively. I do not argue that this approach is *ultimately correct*, but instead argue that it is the best available option in light of our cognitive limitations in effectively formalizing and improving our processes for thinking about infinite ethics and long shots.

## 7.1 The fanatical approach

As we saw in section 6.2.4, Period Independence, one of my key assumptions in arguing for the overwhelming importance of the far future, implies fanaticism. In this section, I explore the consequences of fanaticism for the value of shaping the far future.

### 7.1.1 Some complications in comparing options with infinite expected value

There are many complications with comparing infinite expectations. For instance, some difficult questions to ask about infinite expectations include the following:

1. Do higher orders of infinite value lexically dominate lower orders of infinite value?

2. Among different orders of infinite value, are there gradations of infinite value? For example, does it make sense to say that both $A$ and $B$ are infinitely good and of the same order of infinity, but one is better than the other?

3. How does a gamble whose expected value is infinite, but will definitely have a finitely valuable outcome (such as the St. Petersburg gamble) compare with a gamble that has some probability of producing an infinitely valuable outcome?

4. How do we compute the value of an outcome where there are infinitely many periods of negative value and infinitely many periods of positive value? It is well known that the sum of such an infinite series varies depending on the order in which you compute it. Is there some privileged ordering to use when computing such sums?

5. How do we compare the possibility of getting infinite value along different possible dimensions of value, such as the number of periods and the value per period? For example, how does

having infinitely many periods at a specified level of value compare with having finitely many periods, each of which has infinite value?[1]

There are doubtless many other such questions and problems associated with developing this view which I won't try to answer here, and probably couldn't answer at all. Still, we can say some constructive things about what it would be best to do, given some set of answers to these questions. Somehow or other, fanaticism implies that it is best to effectively ignore all finite considerations, and to optimize for obtaining some level of infinite value, perhaps weighted for how much we have at that level. And then the question is, of the options available to us, what does that best?

### 7.1.2 The problem of saturation

If we answer the above questions in certain ways, it may be very likely that, no matter what we do, we can't make things much better. I call this the *problem of saturation*. As we'll see below, this problem is particularly acute and strange if there are gradations of infinite value within a given order of infinity, but the problem may not be present otherwise.

#### 7.1.2.1 If there are not gradations of infinite value within a given order of infinity

Consider three possibilities, which we will return to over the course of this chapter:

1. *Big Universe Hypothesis*: Our universe is infinitely large, and contains an infinite amount of intelligent life.
2. *Medium Universe Hypothesis*: Our universe contains a large finite number of instances of intelligent life.
3. *Small Universe Hypothesis*: Our universe contains only a few instances of intelligent life.

Which of these hypotheses is true would have profound effects on how much anything our civilization could do would shift the probability of achieving an infinitely valuable outcome, rather than a finitely valuable outcome. If we live in a Big Universe, we will have essentially zero probability of being able

[1] For further discussion of questions 1, 2, and 4, see Bostrom (2011). For a related discussion to question 4 in the decision theory literature, see Nover and Hájek (2004) and many further discussions of their "Pasadena game."

to create an infinitely valuable outcome where we otherwise would not have had one (since so many other groups would independently have this chance). If we live in a Small Universe, we'll have the best shot, since we may be the only ones who can do this. A consequence of this is that if we're uncertain what kind of universe we live in, if our goal is to maximize the probability of an infinitely good outcome (to say nothing of how infinitely good it is), it would be best to act as if we do not live in a Big Universe.

This is a strange result because the type of fanatical approach currently under discussion is fundamentally based on the idea of additive separability, and was motivated by the thought that what is going on in causally disconnected parts of reality could have no bearing on what it would be best to do. (Recall the cases of Infinite Physics Research from the previous chapter, and Our Surprising History from chapter 3.) But we're getting exactly that result if we say don't allow for gradations of infinity. I am therefore inclined to think that this type of "mixed" approach is not very plausible.

I can, however, make some general remarks about how we might be most likely to create an infinitely valuable outcome when we otherwise would not have had one. In order for some person's action to make an outcome infinitely better, three very strange things have to be true: (i) it is possible for the person to create an infinite amount of value, (ii) no one else has already done anything which ensures that there will be an infinite amount of value (of the same order), and (iii) no one else will ever do anything that ensures that there will be an infinite amount of value of the same order (provided the person doesn't take the action). Consider our three categories of future benefits: speeding up development, existential risk reduction, and other trajectory changes. While speeding up development could help people produce a larger amount of something that is infinitely valuable, it is very unlikely to satisfy conditions (ii) and (iii). This would only happen if the limiting factor on achieving infinite value was additional research time and we just barely missed it, which seems unlikely given the potentially enormous amounts of time our descendants might survive. By definition, a small trajectory change could not be the difference between a finite and infinite amount of value; but a very large trajectory change, such as averting an existential catastrophe, could do the trick. Successfully reducing existential risk could, however, have all three of these properties. For (i), infinite value might follow from some yet unknown insight or technological development, (ii) might be true if one were dealing with an existential catastrophe, such as nuclear war, that could not possibly have been foreseen by earlier generations, and (iii) might be true because future people can do nothing about existential catastrophes when the critical actions must take place before they exist.

#### 7.1.2.2 If there are gradations of infinite value within a given order of infinity...

There are two cases to consider here. The question is whether we accept, in addition to Period Independence, independence across large spacetime regions. The difference is that under mere Period Independence, the value of what happens in this period nearby Earth may depend on what is happening in remote regions of space. This could not be true if we have independence across large spacetime regions. I discussed this distinction in a previous chapter.

If we accept Period Independence but not independence across large spacetime regions, then what it would be best to do will depend on whether we live in a Big, Medium, or Small Universe, for reasons analogous to the ones discussed above.

If we accept Period Independence and independence across large spacetime regions, then what it would be best to do would probably not depend on whether we live in a Big, Medium, or Small Universe. Why? Well, the additional value of what we do will not depend on what is happening in these remote parts of the universe, so as long as other civilizations don't interact with us, they can be safely ignored for purposes of decision-making. In finite world, a total utilitarian would not care how many people exist in other parts of the world if he was deciding whether to increase the total well-being by a certain amount; he'd only care by what finite amount he increased the total goodness of the world. Similarly, if we accept the package of views in question, the number of other civilizations in the universe would not affect what it would be best for us to do, since it would not affect by what infinite amount we could make the world better. Thus, it would not matter if we lived in a Big, Medium, or Small Universe. Thus, in this scenario, smaller trajectory changes would not necessarily be unimportant.

## 7.2 Timid approaches, Period Independence, and the value of shaping the far future

What does a timid approach imply about the value of shaping the far future? As we saw in the last chapter, this requires using a bounded utility function to rank alternatives. There are different ways to do this, and they have different implications about the importance of shaping the far future. In this section I discuss the different ways that this could be done and their implications.

### 7.2.1 If we're timid, shall we try to retain the spirit of Period Independence?

The "minimal change" way to introduce upper and lower bounds to the utility function is as follows:

> *Weakened Period Independence*: By and large, at least when the sum of value across periods is not yet enormous, how well history goes as a whole is a function of how well things go during each period of history; when things go better during a period, that makes the history as a whole go better; when things go worse during a period, that makes history as a whole go worse; and the extent to which it makes history as a whole go better or worse is independent of what happens in other such periods. In very extreme cases, good and bad periods have negligible marginal value.

Roughly, that means that when trading off between number of good periods of history and probability of getting them, a very large increase in the number of good periods will not be an improvement if it comes with even a small decrease in the probability of getting the payoff and both alternatives involve a very large number of good periods. In contrast, on the fanatical approach outlined above, decreasing the probability of payoff by 1/2 can always be compensated for by doubling the number of good periods in a payoff. This difference can be illustrated by examining graphs of the two types of utility functions that I have in mind.

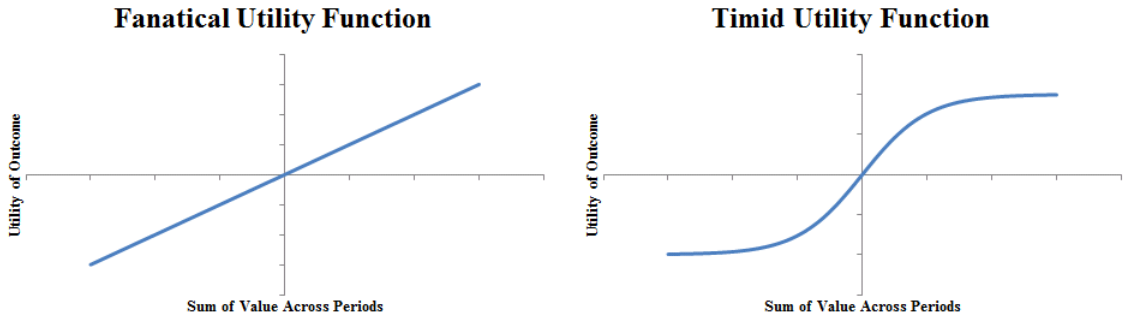


Figure 7.1: Fanatical and Timid utility functions

Something like this is pretty inevitable on the timid approach, provided we want to hold on to approximately ranking histories of the world in terms of the sum of value across different periods of time.

Once we introduce a bounded utility function, we can't be totally true to the intuitions that support Period Independence. Even if we can say that *in practice* the additional utility provided by some additional good periods of history is independent of what has happened in previous periods, we can't accept it *in principle*. We have to admit that, if it really turned out that there had been a lot of great periods in the past, as in Our Surprising History, that could affect what it was best for us to do. And there is some feeling that once we let in some holistic considerations across periods, there is no principled reason to keep us from letting holistic factors playing a larger role.

This feeling can be amplified when we notice structural similarities between the problem of recklessness/fanaticism and problems like the Repugnant Conclusion or the Single-Life Repugnant Conclusion.[2] All of these problems can be viewed in terms of trade-offs between two dimensions, where we make trade-offs between the dimensions in one way in certain cases, and in another way in other cases, and when we put all of that together, we rank outcomes or prospects in a non-transitive way. To see this, consider the following familiar graph:


Figure 7.2: The Repugnant Conclusion and similar problems (repeated)

Now consider four interpretations of this graph. On the standard interpretation, the height of the graph represents the level of well-being in a population, and the width of the graph represents the number of people in the population. On the second interpretation, the height of the graph

[2]For more on the Single Life Repugnant Conclusion, see Temkin (2012, chapter 4).

is the quality of different periods, and the width is the number of such periods. I have discussed this problem in another chapter, but the take-away is that we can get something rather like the Repugnant Conclusion if we accept Period Independence. On the third interpretation, the height of the graph is momentary quality of life and the width is the duration of life. Now we have a very analogous argument for the so-called "Single Life Repugnant Conclusion," according to which it better to have an extremely long life with a very low momentary quality of life than a very long life with a very high level of quality. On the fourth interpretation, height is probability of the payoff, and width is the duration of the payoff at a fixed level of quality (years of life for an individual at a given level of happiness, years that civilization flourishes at a given level of flourishing). On this interpretation, we see a graphical representation of my case of The Devil at Your Deathbed, and an argument for recklessness.

These paradoxes have a common form. Following Temkin (2012), I call them *Spectrum Paradoxes*. The form of the paradox is to claim that

1. Tiny losses along one dimension always yield improvements when they are accompanied by sufficiently large gains along another dimension. (Let's summarize this by saying that $A$-transitions yield improvements.)

2. Something that is very good in terms of two dimensions is always better than something that is *enormously* good along one dimension and close to "neutral" along the other dimension. (Let's summarize this by saying that $B$-transitions are not improvements.)

3. There is some series of outcomes or prospects such that each pair in the series is related in the first way, but the first item in the series is related in the second way to the last item in the series. (Let's summarize this by saying that many $A$-transitions yield a $B$-transition.)[3]

And of course, people try to resolve this paradox by denying one of these three claims, or ranking the outcomes or prospects non-transitively.

I mention all of this to say that someone might believe that all of these paradoxes should be resolved in similar ways, since our inconsistent views about them seem to be generated by similar

[3]Note that I am using "transition" in a metaphorical, rather than temporal sense. It is somewhat more intuitive to think of comparisons in terms of having one object and then trading it for another. This may be impossible in some cases. What we're really trying to say is which of two different outcomes or prospects, compared timelessly, is better than the other. The thing to watch out for here is slipping into comparing outcomes or prospects of the form (have $X$ and switch to $Y$), rather than comparing $X$ and $Y$ directly.

principles of reasoning. In particular, someone might say that if we resolve the timidity/recklessness paradox by denying that tiny losses of probability can always be compensated for by large gains in payoffs, we should resolve the population/well-being paradox by claiming that small losses in well-being can always be compensated for by large gains in total population. Or, more to the point, someone might claim that we should involve the value per period/number of periods paradox by claiming that small losses in value per period cannot be compensated for by large gains in the total number of periods.

I find the above reasoning somewhat compelling, but not decisive. The issues in each version of this paradox are similar but not exactly the same, and people may find different resolutions to different versions of the paradox to be plausible. But more importantly, it is consistent and plausible to hold that, in almost all ordinary cases, $A$-transitions really are improvements, but they aren't improvements in very extreme cases. On this view, giving up on Period Independence completely is like refusing to use a thermometer to measure temperature of soup because the thermometer gives inaccurate readings when it's on the surface of the sun. If we take this approach we might broadly trust Weakened Period Independence in cases where there's no funny business going on. Of course, the difficulty here will be in saying what is and isn't funny business. The cases I'll appeal to later, where I use Weakened Period Independence as a guide, may count as "funny business" or they may not, and saying whether they do will depend on intuitive factors that may be beyond formalization at the moment.

In any case, I can't settle how much we should try to be true to the spirit of Period Independence, so I will discuss the implications for the value of shaping the far future if we try to stay true to the spirit of Period Independence and if we do not.

### 7.2.2 If we don't try to stay true to the spirit of Period Independence...

Views that don't try to stay true to the spirit of Period Independence are diverse and may have little in common, sort of like "views that don't stay true to the spirit of Marxism," so it is impossible to say anything very general about what follows for the value of shaping the far future if adopt one of these views. However, there are some variations that are *similar enough* to the Period Independence approach in the cases that matter, and in these cases our analysis can be similar to the cases where we stay true to the spirit of Period Independence, or learn something important by examining what is best if we are true to the spirit of Period Independence. There are a couple of simple ways this

might be true.

First, if we introduce a more holistic approach to measuring the value of outcomes and prospects, that approach may increase the value of shaping the far future, may be neutral with respect to the importance of shaping the far future, it's effect on the value of shaping the far future may be ambiguous, or its effect on the value of shaping the far future may not be very significant. For example, I argued in chapter 2 that taking account of holistic facts about the distribution of value across periods of time are likely to fall into this category. I considered facts about *averages*, *peaks* (how well things go at the best times), *troughs* (how badly things go at the worst times), our *trajectory* (whether things are getting better or worse), *variety* across periods of time, or *equality* across periods of time. I argued that many of these considerations might fall into one of the above categories (favor shaping the far future, neutral with respect to shaping the far future, unclear significance, or not very important for the value of shaping the far future).

Second, even if Period Independence isn't the *full* story about the value of outcomes or prospects, and even if it isn't approximately the full story about the value of outcomes or prospects, it may be *part* of the full story. It may turn out that Period Independence approximates *one* kind of ideal that is relevant for determining the value of outcomes or prospects. And in this case, it could be very relevant to know what Period Independence implies about the value of shaping the far future.

But I can't say much about how important it would be to shape the far future if something like Period Independence is not even approximately correct in most cases we can think about, or if it were not even approximately correct as one part of the full story about morality. Someone who thought both of these things might help advance the discussion by clarifying their views and explaining what these views imply about the value of shaping the far future.

### 7.2.3 If we try to stay true to the spirit of Period Independence...

In this subsection, I'll discuss what follows if we try to stay true to the spirit of Period Independence. I'll operationalize that by assuming Weakened Period Independence, and discussing what follows given that assumption.

#### 7.2.3.1 If we have an unlimited scope of concern...

Given timidity and Weakened Period Independence, the value of shaping the far future depends somewhat on what I called, in chapter 5, our theory's *scope of concern*. Many theories of the value of outcomes have at least the following three parts:

1. Domain: A domain of objects, properties, and relations that matter

2. Value assignment rule: A rule for saying how good an outcome is if those objects, properties, and relations are in certain conditions

3. Scope of concern: A rule that says which of the objects in the domain are relevant for assigning value to outcomes, or for making goodness-motivated decisions

There's a relatively simple way that our theory's scope of concern bears on the value of shaping the far future, given that we accept timidity and Weakened Period Independence. Given Weakened Period Independence, when the sum of value across periods of history is sufficiently large, the marginal utility of additional periods of history is low. This means that if we have an unrestricted scope of concern, the value of shaping the far future is sensitive to whether we live in a Big, Medium, or Small Universe.

If we live in a Big Universe, then there are infinitely many civilizations, and the sum of value across periods is very likely to be either infinitely positive, infinitely negative, or undefined. In either of the first two options, it will matters very little what we do, so the possibilities should be ignored for practical purposes unless we have overwhelming evidence that we live in a Big Universe. To see why, consider each case.

1. If we're in the case where the sum of value across periods is infinite and positive, then the value of the outcome should already be at or near its upper bound. In that case, we cannot improve things by very much, if at all. We might be able to make things somewhat worse, but again it will be hard to change the value of the outcome by any significant amount because we are only one civilization among many, and the total value of the outcome is very insensitive to differences in the sum of value across periods when the sum of value across periods is already very large.

2. If we're in the case where the sum of value across periods is infinite and negative, then the value of the outcome should already be at or near its lower bound. In that case, we cannot make things much worse than they already are. We may be able to make things somewhat better, but again it will be hard to change the value of the outcome to any significant amount because we are only one civilization among many, and the total value of the outcome is very insensitive to differences in the sum of value across periods when the sum of value across periods is already very small.

To make this a bit more intuitive, you can look at the example timid utility function produced in Section 7.2.1, and note that the effects of changing the sum of value across periods is very small on the far ends of the horizontal axis.

What should we say about the outcome where the sum of value across periods is undefined? That is an extremely challenging open mathematical and philosophical problem, and I'm going to leave it for someone else to deal with. For further discussion of the shape of the problem and some potential solutions, see Bostrom (2011).

So the bottom line is that, given this type of normative system and the background assumption of a Big Universe, increasing the value or number of good periods, or making bad periods good, has essentially no value, or at least isn't the kind of thing we know how to think about yet. Once again, it may seem that something has crashed if living in a Big Universe implies that nothing we do really matters very much. I am inclined to think that there probably is something wrong here, but, as noted in Section 7.1.2.1, there may be something very tricky about infinities that makes this result hard or impossible to avoid.

If we live in a Medium Universe or a Small Universe, then our actions probably matter. If we live in a Small Universe, then our actions might matter more, since we might be closer to the upper or lower bound of how good or bad things could get. (This is analogous to what we found in Section 7.1.2.1.) How much would they matter, and can we think about this type of case in a way that's similar to anything else we've already thought about? That depends on a further question: how high are the bounds on our utility function? I'll consider two possibilities: (i) pick an extremely large upper bound and run with it, and (ii) choose the upper bound in some other way.

#### A very large upper bound

There are various reasons it may be attractive to simply choose a very large upper bound and run with it. A major factor might be the following. Some people, myself included, may not be deeply disturbed by recklessness, but be very disturbed by fanaticism. There is something extremely disturbing about not caring how many zeros there are when someone says that doing action $A$ would give you a .00000...00001 probability of immortality, and doing action $B$ would "only" give you high probability of having a great life for a million years, whereas I don't find it *too* implausible to believe that for any probability $p$, there is some great option $C$, such that getting $C$ with probability $p$ is better than doing action $B$. It's as if things worked out just fine in finite cases, but then infinities came in and mucked everything up. Unfortunately, we can't avoid fanaticism without altering the

reckless approach. So it may seem natural in these circumstances to try to choose an upper bound for our utility function that is similar as possible to the reckless approach, but avoids fanaticism. Inconveniently, there is no such thing as being "as similar as possible to the fanatical approach without reaching it." A timid approach gets more and more similar to a reckless approach as the upper bound on the utility function increases, but you never get very close to "no upper bound at all," just as an increasing sequence of finite numbers can never get very "close" to infinity. But many trade-offs have to be made in ways that feel arbitrary, so perhaps we should just pick a ridiculously large finite number—get a team of smart mathematicians together for a month and see what the biggest number they come up with is—and set the upper bound of our utility function there.

If we take this approach, then the case for shaping the far future would proceed exactly as it did when I naively laid it out in chapter 3, where we left infinite considerations to the side. We can just pick this upper bound large enough that, provided we live in a Small or Medium Universe, we can mostly ignore infinite considerations and just do what would be best with respect to finite considerations.[4]

This approach feels fairly unsatisfying for a couple more reasons. The approach is motivated by trying to make problems of infinity go away,[5] and the methods that are being used to do it are extremely crude. One would expect the truth about this issue to be neater, more elegant. The relatively crude methods that we might use to avoid these problems seem likely to introduce unforeseen complications elsewhere, much like a a poorly patched piece of software. The second worrisome issue is that choosing a very large upper bound means we have to continue to deal with many of the problems of recklessness. And there's a natural complaint about this. The complaint goes like this:

> Look, we're already biting some bullets by accepting timidity. In particular, we're saying that in certain cases, we wouldn't sacrifice a very small amount of probability of getting a big payoff in order to make the payoff *much, much* larger. It may seem that it isn't

[4] We would only "mostly" ignore infinite considerations because it seems pretty unlikely that we will be able to achieve infinite value. Some people might be inclined, given a bounded utility function, to set the value of an infinite sum of value across periods equal to the upper limit of how good finitely good outcomes could be. While this is mathematically elegant, it seems implausible because it implies that a very large sum of value across periods is almost exactly as good as infinitely large sum of value across periods. Provided it could be done consistently, and it seems it can, it would be better to make the infinitely good outcome better than this by some significant multiple. To put it in very plain language, we probably should say that a chance of having a great outcome forever is much better than a chance of having a good outcome for a very long time, and we probably could set up our technical apparatus so that we could say this consistently, despite the fact that another choice might be more mathematically elegant.

[5] Note that this type of approach cannot obviate the need for dealing with infinities, it can only stop them from dominating our lives. We still have to say something about how much value to assign to outcomes with infinite sums of value across periods, including the nasty cases where the order of the sum effects the result you get.

> much worse to say that these cases happen a bit more frequently than you might think, provided it doesn't affect most people's ordinary lives by that much. That way, we can at least avoid taking some more of those crazy long shots that reckless people take!

There's something pretty reasonable about this complaint, and I discuss some ways we could get around these difficulties in the next section.

**Some other upper bound**

I can't say very much in general about how a plausible upper bound could be chosen. In principle, I would recommend using the upper bound that was the output of some reflective equilibrium process that took due care to avoid some of the biases that we discussed in "How Could We so Wrong?" But that isn't very helpful as a piece of practical ethics.

Having a low upper bound decreases the value of extreme long shots, and therefore decreases the value of shaping the far future. These effects are much more pronounced if we live in a Small Universe. In that case, the crucial question is how many times better an outcome we could get is than the outcome we've got, assuming our future went very well. With an unbounded utility function, the answer might be "trillions of times better," but the answer could be much lower if we set the upper bound of our utility function differently. Our answer greatly affects the expected value of existential risk reduction. I can't say anything terribly useful about how high it would have to be set in order to set back the case for shaping the far future, but if it is much lower, then that might decrease the value of shaping the far future.

If we live in a Medium Universe, it is less clear what the implications are for shaping the far future. If it is a big enough Medium Universe, then there will already be quite a lot of value across all periods, and putting an upper bound on the utility function will also reduce the value of many other things we do. So it is less clear what we should say about existential risk in this case. If we are in a very large Medium Universe, then we may already be very close to the upper bound of our utility function, in which case what we do would matter very little, and it would make sense to ignore the possibility for practical purposes, as long as the evidence that we were in that scenario were not very strong.

Either way, there is a significant complication. Rather than reducing existential risk, we may try to shape the far future by creating other positive trajectory changes or speeding up humanity's

development. But in scenarios where the future goes very well, the outcome may be so good that it is close to the upper limit of how good things could be. In that kind of case, small trajectory changes or speeding up development may have essentially negligible value. Taking account of all these issues is rather messy, and makes it very difficult to speak confidently about the comparative value of different methods of shaping the far future, given a bounded utility function with an unknown upper bound.

#### 7.2.3.2 If we have a limited scope of concern...

As we saw in section 7.2.3.1, having an unlimited scope of concern and combining it with a bounded utility function introduces some rather strange conclusions. In particular, in Big Universe scenarios, it matters very little what we do, though we could positively or negatively affect the lives of many people.

One solution to this problem would be to limit our scope of concern in such a way that what it would be best for us to do would not depend on what happens in parts of the world over which we have no control. When deciding how to limit the scope of our concern, we want to be very careful not to limit the scope of our concern in such a way that there is nothing that intuitively matters, is under our control, and claimed to be outside of our scope of concern. We wouldn't want to say that events outside our light cone are beyond our scope of concern and then discover that there is some new technology that allows us to control events outside of our light cone—that would be pretty embarrassing! So we should choose a scope of concern so that events within our control are always part of our scope of concern.

What would the implications be for the importance of shaping the far future if the proper scope of concern were limited to the events under our control? In terms of what it would be best to do, it would be essentially equivalent to a case in which we lived in a Small Universe, since we could effectively ignore all the Big Universe stuff that would be beyond our control, and lots of our future would be under our control.[6] There may be some differences because different people would have different aspects of the future under their control, but it seems that the analysis would proceed roughly in the same way as before.

[6] There are some concerns remaining here. On some plausible decision theories, it may be that what our counterparts do counts as "under our control" in the decision-relevant sense. For example, consider the fact that if we live in a Big Universe, then there are infinitely many people that are essentially exactly like you and I dispersed throughout the universe. Some of them will be doing things that are essentially exactly the same as us. I should believe that if I face a choice between $A$ and $B$, and I choose $B$, then many of these counterparts will also choose $B$. Therefore, on evidential decision theory, the effects that get counted when I calculate the impact of choosing $A$ or $B$ will be infinite, since I'll have infinitely many counterparts.

## 7.3 How to get by until we have better answers

We could summarize the situation as follows. In chapter 3, I defended a plausible set of views which suggest that shaping the far future is overwhelmingly important. This set of views has plausible implications in a wide set of circumstances, but it has implausible implications about long shots and infinite ethics. The main way of altering these views—what I called the timid approach—has plausible implications for comparing long shots considerations and ordinary considerations, and has plausible implications in a wide variety of ordinary cases. However, it has implausible implications about infinite cases and some extreme cases. It is far from clear which of these two types of views is most plausible, and they may have different implications about the value of shaping the far future. Yet, we may have to make choices about shaping the far future before we can solve these challenges. So the question is: what should we do until then?

One approach, which I'll call the "methodological monist" approach, would suggest that we should select one theory (or kind of theory) that we have considered and apply it everywhere. We could select one specific view (e.g., timidity + weakened period independence + limited scope of concern) and use it to answer all questions about the importance of shaping the far future. Another approach, which you could call the "methodological pluralist" approach, would be to select different theories to use in different circumstances. This is similar to what Temkin (2012) calls an *essentially comparative view*, except methodological pluralists do not necessarily claim that the non-transitive rankings yielded by their approach are *true*. Methodological pluralists claim only that their approach is the most reasonable way to proceed at the moment. And finally, of course, we could adopt an "intuition-based" approach where we claim that all of our theories are unacceptable and we rely mainly on our (inconsistent!) intuitions to make decisions in cases that involve existential risk.[7]

The main claims I'd like to make on this issue are as follows:

1. For reasons I articulated in chapter 2, the paradoxes of infinite ethics and long shots do not provide compelling reasons to follow the intuition-based approach.

2. All of the theories we have considered introduce implausible, unintended consequences in certain types of cases, and this makes the methodological monist approach unsatisfying.

3. Consistency-based arguments do not provide compelling objections to the methodological methodological pluralist approach.

[7] It's worth acknowledging that there is a range of intermediate cases between extreme methodological monists and extreme methodological pluralists.

4. In practice, we should try to ignore the fanaticism problem as much as possible and rely on the assumptions I defended in the first part of this dissertation: Period Independence, Additionality, Temporal Neutrality, and expected utility theory. We should provisionally assume that whatever empowers future generations most is also the best bet in terms of infinite considerations on the grounds that if there are opportunities to produce infinite value, we best exploit those opportunities (in expectation) if our descendants survive and thrive long enough to exploit them. If we are forced to reject this assumption, we should use a timid approach to decide whether or not to prioritize infinite considerations.

The approach I defend may not be ideal. I expect that more rational versions of us would not use it. However, in light of our cognitive limitations and the errors that we are likely to make when thinking about the far future, long shots, and infinite value, I'll argue that it is our best bet.

### 7.3.1 Against the intuition-based approach

Someone might argue that in light of the paradoxes we have seen involving long shots, we should rely on an intuition-based approach. This might have been a plausible argument if the paradoxes we've seen cast doubt on the reliability of theoretical approaches but not intuition-based approaches. But this is not the case: these paradoxes show that our intuitions are inconsistent. As I argued at length in chapter 2, there are many reasons to expect our moral judgments in general to be unreliable, and these reasons are especially strong in the case of shaping the far future. When we must form a view on the basis of many unreliable inputs, our best strategy is to try to look for general views that capture most of the big picture and satisfy theoretical virtues, even if those general views conflict with much of our data.

### 7.3.2 Against the methodological monist approach

Some would point to the impossibility results in population ethics, as well as the paradoxes involving long shots that I've presented here, as evidence that the methodological monist approach will be unsatisfactory. I reject this argument. The impossibility results show that our deepest convictions about population ethics are inconsistent, and therefore that no consistent theory can capture all of our deepest convictions. If no consistent view will satisfy all of our deepest convictions, it hardly seems fair to complain that the methodological monism won't meet this standard.

Instead, what makes me wary of the methodological monist approach is that every theory sketch I've seen or proposed myself (including common sense) seems "buggy" in at least some cases. When

I say that an approach is "buggy," I mean that it involves, or looks like it would involve, implausible unintended consequences when applied in certain cases. As I said in chapter 2, there is an important difference between a counterintuitive result and a result that fails to formalize a thought as intended. In software, we can often distinguish between surprising but reasonable output and buggy output because we can have good external checks on whether some piece of software is giving intended results, and we can alter our software in response to that feedback. In trying to say how important shaping the far future is, as with normative theory in general, the challenge is that we don't have a good idea of what the "intended result" is and we don't have many external checks to rely on, so it is hard to learn from mistakes.

However, to continue with the programming analogy, there are some conditions under which we can expect a program to be more likely to behave in a buggy way, even if we don't have an easy way to confirm that the program is yielding unintended results. A program would be especially likely to be buggy if:

1. The program is being used in a novel domain, especially a domain where the developers had not thoroughly tested the program, a domain where the programmers had not specifically intended the program to work, or a domain where other programs have not been used extensively before. (For example, suppose a program divides real numbers but is primarily used to divide positive integers. Negative integers, fractions, and irrational numbers would count as novel domains and be more likely to produce buggy output.)

2. The program was being run in an "edge case," meaning a case where one of the inputs to the program took an extreme value. (In our division program, edge cases might include zero, infinity, very small numbers, and very large numbers.)

3. The program produced errors in one domain, and the programmers altered the program in an inelegant way that avoids the problem in that particular domain, but the programmers did not have a deep understanding exactly what it was about that domain that caused the problem, whether there could be a more general issue that caused the problem, and whether the patch may introduce new problems. (Software people would call these alterations "kludgy.")

Likewise, a normative theory is especially likely to fail if we use it to deliver results in a novel domain or in "edge cases," or if the theory has been patched to avoid certain types of intuitive counterexamples when we are not very sure that we have gotten to the bottom of why the theory generates that type of counterexample.

When we develop theories of population ethics, infinite ethics, and extreme long shots, we face all of these difficulties. Our values and norms have not been stress-tested in contexts that involve existential risk, infinite values, and extreme long shots. Decisions in cases of extreme long shots that involve the far future and infinities are far from the paradigm cases that are typically used to develop normative theories, and even if we focus on these cases in particular, it is hard to make progress because we are dealing with edge cases and a novel domain.

It also seems that many of the solutions to the problems we face are kludgy. When we try to avoid a problem like fanaticism by introducing upper and lower limits to how good outcomes could be, we introduce unintended problems such as extreme risk aversion in some cases, extreme risk seeking in others, and strange violations of Period Independence. When we try to avoid a problem like the Repugnant Conclusion by introducing Person-Affecting Views, critical-level views, or theories of diminishing marginal value of population, we introduce counterintuitive implications about the morality of having children, the Sadistic Conclusion, and various types of risk aversion. All these views can be seen as kludgy alterations of the classical utilitarian approach.

### 7.3.3 In defense of methodological pluralism

We should rely primarily on a methodological pluralist approach, relying on different theories in different contexts. Why? I argue that it is the pragmatically best response to the problems we face, in light of our limitations in effectively thinking about the problems at hand.

The natural complaint about this approach is that it is inconsistent, and it is. Hopefully, that means that it is possible, in principle, to do better. But it doesn't mean that *we* can do better in any practically meaningful sense, and it therefore isn't a good objection to methodological pluralism. A few examples illustrate this. Temkin (2012, p. 504) points out that Niels Bohr's model of the atom was known to be internally inconsistent, but was the dominant model for more than a decade because it had more predictive and explanatory power than any of the alternatives. There's a similar story for Cantorian set theory. Cantor's approach dominated mathematical study of set theory at the end of the 19th century, and it continued to do so after Russell, Zermelo, and Cantor had proven that Cantor's theory was inconsistent between 1899 and 1903. Zermelo developed the first axiomatic approach to set theory in 1908, but mathematicians did not stop using set theory in the interim. It seems clear that, in the absence of an approach that was good enough along other dimensions, such as predictive and explanatory power, it was eminently reasonable for these physicists and mathematicians to continue to use the inconsistent theories that they had. The

reason this didn't lead to disaster was that people using inconsistent theories can be careful to avoid reasoning their way into nonsense, even if an unsophisticated automated reasoning machine could not. For a third, well-worn example, we can consider the fact that quantum mechanics and general relativity are inconsistent with each other, but physicists routinely use both in the contexts where they are confident that the theories work. For a final example, imagine that we discover a difficult-to-resolve inconsistency in the American legal code; I'm sure there is one. We would not conclude, on this basis, that any other consistent and basically decent legal code was superior to the American legal code. Instead, we would (rightly) continue to rely on the American legal code as it stood until the legal code was altered in a way that removed the inconsistency without too great a cost. The lesson here is that while inconsistency may be the final word when it comes to *truth*, it is not the final word when it comes to *practice*.

The foregoing considerations show that it can be rational to accept and reason under a set of assumptions which is known to be inconsistent when it is done tastefully and the alternatives are not sufficiently good. Is it plausible that we are in an analogous situation with respect to the fanatical and timid approaches to dealing with long shots that we have been considering? Yes. These examples from physics, mathematics, and the law all involve a complex practice which was built up for use in certain types of situations where the practice performed well. In each case, it is very challenging to come up with a new system that gets things right in all cases, in large part because there are potentially many ways of proceeding and many of them have not been adequately considered and tested. If we simply "make up" some consistent approach, it seems likely to suffer from bugs that would be difficult to deal with, especially since the people using the new approach could not rely on accumulated wisdom about when the new approach was likely to be unreliable. Analogously, the approaches a methodological pluralist might rely upon arise from a complex practice which relies on different techniques in different cases, which perform well in different domains. All known ways of uniting these approaches into one consistent approach introduce bugs that we don't know how to handle without deviating from any unitary approach.

A specific version of the inconsistency complaint is the "money pump" argument. According to this argument, an agent with inconsistent preferences can start with a certain set of goods and make a series of trades, each of which he regards as an improvement, and leaves him worse off than he was when he started. For instance, if the agent prefers $A$ to $B$ and $B$ to $C$ and $C$ to $A$, the agent might start with $A$ and then pay to move to $B$, pay to move to $C$, and then pay again to move back to $A$. As Temkin (2012) and others have shown, people do have inconsistent preferences with this kind of structure, and I have argued that people have inconsistent preferences about extreme long shots.

Yet, we never hear about people who get money pumped. Why? One possibility is that people never get offered these trades that would trigger money pumps. A more plausible answer is that people do not act on their preferences in the inflexible way this argument assumes. When they get into a situation where they see that their preferences would lead them to get money pumped, they either change their preferences or refuse to continue to act on some of those preferences. Because of this, money pump arguments do not illustrate a practical danger for humans. It is plausible that having preferences which would be theoretically susceptible to a money pump displays a failure of perfect rationality, but, once again, that a certain approach is imperfect does not imply that an improved approach is meaningfully available.

### 7.3.4 The approach I favor

The specific methodological pluralist approach I favor can be summarized as follows:

> When comparing finite outcomes, use the approach I developed in the first half of the dissertation (Additionality, Period Independence, Temporal Neutrality, and expected utility theory). Assume, in general, that whatever is best for shaping the far future is best with respect to infinite considerations. If this assumption seems to be mistaken and you must compare infinite considerations and finite considerations, follow a timid approach.

Why favor this approach? In short, the reasons are that the theory I developed in the first half of the dissertation seems reasonable in many finite cases but likely to be buggy when comparing infinite cases with finite cases; the timid approach seems reasonable and less likely to be buggy when comparing infinite cases and finite cases; and, as I argued in section 6.4.3, future people seem to be in a much better position than us to achieve infinite value, if it is possible at all. I elaborate on the first two points below.

Let's call the approach I developed in the first half of the dissertation the *Basic Approach*. Chapters 2-5 argue that this approach is reasonable in finite cases. Why expect the Basic Approach to be buggy in infinite cases? As we have already seen, this approach implies fanaticism if followed strictly in both infinite and finite cases. But the approach was developed primarily with finite cases in mind, so infinite cases are edge cases in a novel domain for the theory. Moreover, the assumptions of expected utility theory (the continuity axiom, in particular) break down once we start considering infinite cases, and ways of developing the theory further seem kludgy. We just have very little reason to expect the Basic Approach to deliver reasonable results in infinite cases, and intuitively, it doesn't

seem to be working very well.

What about using timid approaches to decide between finite and infinite considerations? Should we expect more or fewer errors using this approach rather than the Basic Approach? The timid approach is better in one way because the approach was at least designed to deliver non-crazy results in long shots and infinite cases, so at least the theory is being applied in its central domain. But we are still dealing with edge cases. And the timid approach is somewhat kludgy; there is no elegant way to select an upper limit for how good outcomes can be, the upper limit was introduced to solve a certain type of problem that arose for reasons that were not originally anticipated, and the change created new problems that were not anticipated. However, the most severe and buggy-seeming of these problems are the violations of Period Independence and the conclusion that if there is a Big Universe, consequences don't matter very much. These distortions are somewhat independent of the distortions introduced by fanaticism, so it may be that the kludge did not create problems in the central domain for which the timid approach was designed. And, of course, the timid approach seems to have more intuitively acceptable implications about making tradeoffs between infinite and finite considerations. So there are some reasons to be more comfortable with the implications of timid approaches when comparing finite and infinite considerations.

## 7.4 Conclusion

In this chapter, I considered various combinations of views about fanaticism, Period Independence, whether we accept gradations among infinities, whether we live in a Big Universe, and whether we have a limited or unlimited scope of concern. These combinations of views had various consequences for the value of shaping the far future, and all had some very unpalatable implications.

In light of these challenges, I advocated that we follow the fanatical approach in the cases where it has plausible implications, and the timid approach in the cases where it has plausible implications. I claimed that this approach is likely to be superior to pure reliance on intuition for the reasons I articulated in chapter 2. Though the approach I defended relies on an inconsistent set of assumptions, I argued that it can be rational to accept and reason under a set of assumptions which is known to be inconsistent when it is done tastefully and the alternatives are not sufficiently good, and that this exception applies in this particular case. I did not argue that this mixed strategy is ultimately correct, but instead that it is the best available option in light of our cognitive limitations in effectively formalizing and improving our processes for thinking about infinite ethics and long shots. Applying this mixed strategy supports my thesis that shaping the far future is

overwhelmingly important.

# Chapter 8

# Conclusion

I'll close by summarizing the course of investigation in this dissertation. Having covered a lot of material and developed language for talking about the problems at hand, this summary is somewhat different from what I presented in the introduction. But the remaining questions are basically the same, so the reader should look at chapter 1 for my views about what related questions deserve additional attention.

## 8.1 Developing the case for the overwhelming importance of shaping the far future

The aim of this dissertation was to evaluate the idea that shaping the far future is overwhelmingly important. To do that, I started with what I called "the rough future-shaping argument." That argument goes as follows:

1. Humanity may survive for millions, billions, or trillions of years.
2. If humanity may survive may survive for millions, billions, or trillions of years, then the expected value of the future is astronomically great.
3. Some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in not-ridiculously-small ways.
4. If the expected value of the future is astronomically great and some of the actions humanity could take would be expected to shape the trajectory along which our descendants develop in

not-ridiculously-small ways, then from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

5. Therefore, from a global perspective, what matters most (in expectation) is that we do what is best (in expectation) for the general trajectory along which our descendants develop over the coming millions, billions, and trillions of years.

I focused on assessing the normative assumptions relevant to premises 2 and 4.

In defense of premises 2 and 4, I argued that it would be good for there to be additional future generations (by appeal to Additionality), that the value of additional future generations did not diminish as the number of past generations increased (by appeal to Period Independence), that we should be neutral with respect to time (Temporal Neutrality), and that expected utility theory is the right tool for assessing the value of shaping the far future. I argued that these assumptions, together with additional empirical assumptions, made the rough future-shaping argument plausible. I defended this argument from attacks based on Person-Affecting Views and views according to which additional lives, generations, or good periods of history have diminishing marginal value.

## 8.2 The challenge of fanaticism

Having defended the major assumptions of population ethics that support the idea that shaping the far future is overwhelmingly important, I explored a challenge to the package of views that I defended. The challenge was that these views implied fanaticism: the view that what it would be best to do would almost always depend on what would be best with respect to infinite considerations. On this view, what it would be best to do turns on questions like:

- What would maximize the probability of an infinitely good future?
- What would minimize the probability of an infinitely bad future?
- What would create the "greatest infinite expectation" of value?

Rather than more normal questions, like:

- Which of these global health programs would save the most lives?
- Which of these policies would maximize GDP?
- Which of these lesson plans would maximize student achievement?

We got into this mess because Period Independence, Additionality, and Temporal Neutrality imply that there is no upper limit to how good outcomes can be. And if there is no upper limit to how good outcomes can be, I argued, then maximizing expected good requires fanaticism.

I also argued that fanaticism is very hard to avoid. Fanaticism cannot be avoided unless there are limits to both how bad and how good outcomes can be. And if there are such limits, then we must accept a "timid" theory. On these theories:

- Sometimes, one prospect is worse than another even though the payoff is *vastly* better, and the probability of getting the payoff only *very slightly* lower.
- If things are going to be very bad anyway, it is rational to be extremely risk-seeking.
- If things went very badly or very well in the distant past, this could have very significant implications about what it would be best for us to do.

Worse still, if we have any uncertainty about whether there is an upper limit to how good outcomes can be, there is a sense in which fanaticism is "inherited" under moral uncertainty. If we decide to give some weight all the moral theories in which we have some credence, we follow expected utility theory, and we say that one action is preferable to another (under moral uncertainty) if it is preferable on some theories and worse according to none, then if we have *any* credence in fanatical theories, our procedure for making decisions under moral uncertainty must also be fanatical.

In the previous chapter, I further considered the consequences of accepting a fanatical theory or a timid theory. On fanatical theories, what it would be best to do depends on whether there are gradations within different levels of infinite value (so we can say that two outcomes are both infinitely good, but one is better than the other), whether or not we live in a Big Universe, and whether how good it would be if things go well for us metaphysically depends on what is happening in distant regions of space. Depending on the answers to these questions, it could be that it would be best to maximize the probability of an infinitely good future, best to do something like "maximize infinite expected value," or the consequences of what we do may matter very little—all strange and surprising conclusions.

On timid theories, what it would be best to do depends on whether we accept a weakened form of Period Independence, how large the upper limit is on how good outcomes could be, whether we limit what I called our "scope of concern," and whether we live in a Big Universe. Depending on what we say about these questions, it could be that it would be best to do roughly what we would be best given the Basic Approach I defended in the first part of the dissertation; it could be that the

consequences of what we do may matter little; or, if the upper limit is small, it may be best to do good in more ordinary ways and for more conventional reasons; or, if we do not accept a weakened form of Period Independence, saying what it would be best to do would may require a very different kind of analysis than anything I have provided here.

Since choosing an adequate fanatical or timid theory poses very great challenges, I argued that, in practice, it would be best to rely on each kind of theory in the cases where that kind of theory has the most plausible results. This can lead to using an inconsistent set of assumptions, but it can be rational to accept and reason under a set of assumptions which is known to be inconsistent when it is done tastefully and the alternatives are not sufficiently good. This exception, I argued, applies in this particular case. Applying this mixed strategy, I argued, supports my original thesis that shaping the far future is overwhelmingly important.

## 8.3 Final remarks

I've argued that shaping the far future is overwhelmingly important. There is significant normative and empirical uncertainty about how best to do this. Resolving this uncertainty would be very valuable, since it may help us to understand what matters most. All things considered, rather little effort has been directed toward this purpose and the issues do not seem completely intractable, so it is not irrational to hope that we can resolve some of this uncertainty.

# Bibliography


Adams, F. C. (2008). Long-term astrophysical processes. In Bostrom, N. and Cirkovic, M. M., editors, *Global Catastrophic Risks*, pages 33–47. Oxford University Press.

Alpert, M. and Raiffa, H. (1982). A progress report on the training of probability assessors. *Judgment under Uncertainty: Heuristics and Biases*, pages 294–305.

Arrhenius, G. (2000). *Future Generations: A Challenge for Moral Theory*. PhD thesis, Uppsala.

Arrhenius, G. (forthcoming 2013). *Population Ethics: The Challenge of Future Generations*. Oxford University Press.

Arrow, K. J. (1971). *Essays in the Theory of Risk Bearing*, volume 40.

Baker, A. (2011). Simplicity. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Summer 2011 edition.

Baron, J. and Greene, J. (1996). Determinants of insensitivity to quantity in the evaluation of public goods. *Journal of Experimental Psychology: Applied*, 2:107–125.

Blackorby, C., Bossert, W., and Donaldson, D. (2005). *Population Issues in Social-Choice Theory, Economics, and Ethics*. Cambridge University Press.

Bostrom, N. (2002). Existential risks. *Journal of Evolution and Technology*, 9.

Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, 15(03):308–314.

Bostrom, N. (2009). Pascal's mugging. *Analysis*, 69(3).

Bostrom, N. (2011). Infinite ethics. *Analysis and Metaphysics*, 10:9–59.

Bostrom, N. (forthcoming 2012). Existential risk prevention as global priority. *Global Policy*.

Bostrom, N. and Ćirković, M. (2008). *Global Catastrophic Risks*. Oxford University Press, USA.

Broome, J. (1991). *Weighing Goods: Equality, Uncertainty, and Time*. Basil Blackwell.

Broome, J. (1992). *Counting the Cost of Global Warming*. White Horse Press.

Broome, J. (1999). *Ethics Out of Economics*. Cambridge University Press.

Broome, J. (2004). *Weighing Lives*. Oxford University Press.

Broome, J. (2008). The ethics of climate change. *Scientific American*, pages 69–73.

Broome, J. (2010). The most important thing about climate change. In Boston, J., Bradstock, A., and Eng, D., editors, *Public Policy: Why Ethics Matters*, pages 101–116. ANU E Press.

Carnap, R. (1950). Logical foundations of probability.

Chapman, C. (2004). The hazard of near-earth asteriod impacts on earth. *Earth and Planetary Science Letters*, 222:1–15.

Cowen, T. and Parfit, D. (1992). *Justie Between Age Groups and Generations*, chapter Against the Social Discount Rate, pages 144–161. Yale University Press, New Haven.

Daniel Kahneman, P. S. and Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge Univ Press.

Erler, A. (2010). Should we rid the world of carnivores if we could? http://blog.practicalethics.ox.ac.uk/2010/09/should-we-rid-the-world-of-carnivores-if-we-could/. Accessed October 1, 2012. Archived by WebCite® at http://www.webcitation.org/6B60gNE3G.

Fetherstonhaugh, D., Slovic, P., Johnson, S., and Friedrich, J. (1997). Insensitivity to the value of human life: A study of psychophysical numbing. *Journal of Risk and Uncertainty*, 14(3):283–300.

Fischhoff, B. and Beyth, R. (1975). I knew it would happen: Remembered probabilities of once-future things. *Organizational Behavior and Human Performance*, 13(1):1–16.

Garber, D. (1983). Old evidence and logical omniscience in bayesian confirmation theory. *Minnesota Studies in the Philosophy of Science*, 10:99–131.

Greene, J. and Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, 6(12):517–523.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, 108(4):814.

Hajek, A. (2003). Waging war on pascal's wager. *Philosophical Review*, 112(1):27–56.

Hanson, R. (2002). Why health is not special: Errors in evolved bioethics intuitions. *Social Philosophy and Policy*, 19(2):153–179.

Harsanyi, J. C. (1955). Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63(4):309.

Holtug, N. (2004). Person-affecting moralities. In Ryberg, J. and Tännsjö, T., editors, *The Repugnant Conclusion*, pages 129–162. Kluwer Academic Publishers.

Hurka, T. (1983). Value and population size. *Ethics*, 93(3):496–507.

Jamison, D. T. et al. (2006). *Disease Control Priorities in Developing Countries*. The World Bank and Oxford University Press, 2 edition.

Jones-Lee, M., Loomes, G., and Philips, P. (1995). Valuing the prevention of non-fatal road injuries: Contingent valuation vs. standard gambles. *Oxford Economic Papers*, pages 676–695.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kavka, G. (1982). The paradox of future individuals. *Philosophy & Public Affairs*, pages 93–112.

Knobe, J., Olum, K., and Vilenkin, A. (2006). Philosophical implications of inflationary cosmology. *The British Journal for the Philosophy of Science*, 57(1):47.

Kogut, T. and Ritov, I. (2005). The identified victim effect: an identified group, or just a single individual? *Journal of Behavioral Decision Making*, 18(3):157–167.

Kunreuther, H., Novemsky, N., and Kahneman, D. (2001). Making low probabilities useful. *Journal of Risk and Uncertainty*, 23(2):103–120.

Lockhart, T. (2000). *Moral Uncertainty and its Consequences*. Oxford University Press, USA.

McMahan, J. (2010). The meat eaters. http://opinionator.blogs.nytimes.com/2010/09/19/the-meat-eaters/. Accessed October 1, 2012. Archived by WebCite® at http://www.webcitation.org/6B60s87Nd.

Meacham, C. J. G. (2012). Person-affecting views and saturating counterpart relations. *Philosophical Studies*, 158(2):257–287.

Nagel, T. (1971). The absurd. *The Journal of Philosophy*, 68(20):716–727.

Nagel, T. (1991). *Mortal Questions*. Cambridge University Press.

Narveson, J. (1967). Utilitarianism and new generations. *Mind*, 76(301):62–72.

NASA (2007). Near-earth object survey and deflection analysis of alternatives report to congress. Technical report, NASA Office of Program Analysis and Evaluation.

Ng, Y. (1989). What should we do about future generations? *Economics and Philosophy*, 5(02):235–253.

Nover, H. and Hájek, A. (2004). Vexing expectations. *Mind*, 113(450):237–249.

Parfit, D. (1984). *Reasons and Persons*. Oxford University Press.

Parfit, D. (1997). Equality and priority. *Ratio*, 10(3):202–221.

Parfit, D. (2011). *On What Matters*. Oxford University Press.

Pierson, P. (2000). Increasing returns, path dependence, and the study of politics. *American political science review*, pages 251–267.

Pinker, S. (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Penguin Books.

Posner, R. (2004). *Catastrophe: Risk and Response*. Oxford University Press.

Pronin, E. and Kugler, M. (2007). Valuing thoughts, ignoring behavior: The introspection illusion as a source of the bias blind spot. *Journal of Experimental Social Psychology*, 43(4):565–578.

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Roberts, M. A. (2003). Can it ever be better never to have existed at all? Person-based consequentialism and a new repugnant conclusion. *Journal of Applied Philosophy*, 20(2):159–185.

Ross, J. (2006a). *Acceptance and Practical Reason*. Rutgers University Ph. D. dissertation.

Ross, J. (2006b). Rejecting ethical deflationism. *Ethics*, 116(4):742–768.

Samuelson, P. (1977). St. Petersburg paradoxes: Defanged, dissected, and historically described. *Journal of Economic Literature*, pages 24–55.

Sandberg, A. and Bostrom, N. (2008). Global catastrophic risks survey. Technical report, Future of Humanity Institute.

Scheffler, S. (1994). *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*. Oxford University Press.

Schopenhauer, A. (1942). On the suffering of the world. In Saunders, T. B., editor, *Complete Essays of Schopenhauer*. Willey Book Company.

Schwitzgebel, E. (2010). Kant on killing bastards, on masturbation, on wives and servants, on organ donation, homosexuality, and tyrants. http://schwitzsplinters.blogspot.com/2010/03/kant-on-killing-bastards-on.html. Accessed October 1, 2012. Archived by WebCite® at http://www.webcitation.org/6B5xLAj1N.

Schwitzgebel, E. and Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language*, 27:135–153.

Sepielli, A. (2010). *Along an Imperfectly-lighted Path: Practical Rationality and Normative Uncertainty*. PhD thesis, Rutgers University-Graduate School-New Brunswick.

Sider, T. (1991). Might Theory X be a theory of diminishing marginal value? *Analysis*, 51:265–271.

Sikora, R. (1978). Is it wrong to prevent the existence of future generations? In Sikora, R. I. and Barry, B., editors, *Obligations to Future Generations*, pages 112–166. Philadelphia: The White Horse Press.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy and Public Affairs*, 1(3):229–243.

Singer, P. (2009). *The Life You Can Save: Acting Now to End World Poverty*. Random House Inc.

Slovic, P. (2007). If I look at the mass I will never act: Psychic numbing and genocide. *Judgment and Decision Making*, 2(2):79–95.

Small, D. and Loewenstein, G. (2003). Helping a victim or helping the victim: Altruism and identifiability. *Journal of Risk and Uncertainty*, 26(1):5–16.

Sunstein, C. (2002). Probability neglect: Emotions, worst cases, and law. *The Yale Law Journal*, 112(1):61–107.

Swinburne, R. and Ebrary, I. (1997). *Simplicity as evidence of truth*. Marquette University Press Milwaukee, Wisconsin:.

Temkin, L. S. (1987). Intransitivity and the mere addition paradox. *Philosophy and Public Affairs*, 16(2):138–187.

Temkin, L. S. (1996). A continuum argument for intransitivity. *Philosophy and Public Affairs*, 25(3):175–210.

Temkin, L. S. (1997). Rethinking the good, moral ideals, and the nature of practical reasoning. In Dancy, J., editor, *Reading Parfit*, pages 290–344. Basil Blackwell.

Temkin, L. S. (2000). Equality, priority, and the levelling down objection. In Clayton, M. and Williams, A., editors, *The Ideal of Equality*, pages 126–161. Macmillan and St. Martin's Press.

Temkin, L. S. (2008). Is living longer living better? *Journal of Applied Philosophy*, 25(3):193–210.

Temkin, L. S. (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford University Press.

Thaler, R. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters*, 8(3):201–207.

Unger, P. K. (1996). *Living High and Letting Die: Our Illusion of Innocence*. Oxford University Press.

Velleman, J. D. (2000). Well-being and time. In *Possibility of Practical Reason*. Oxford University Press.

Weitzman, M. (2009). On modeling and interpreting the economics of catastrophic climate change. *The Review of Economics and Statistics*, 91(1):1–19.

Wikipedia (2013). List of cognitive biases. Accessed February 26, 2013. Archived by WebCite® at http://www.webcitation.org/6EiwYMl6Y.

Williams, B. (1973). The makropulos case: reflections on the tedium of immortality. *Problems of the Self*, pages 82–100.

Williams, B. (2006). *Philosophy as a Humanistic Discipline*. Princeton University Press.

Yudkowsky, E. (2008). Cognitive biases potentially affecting judgment of global risks. In *Global Catastrophic Risks*, pages 91–119.