Evaluating Strong Longtermism

Karri Lassi Heikkinen

UCL

Master of Philosophical Studies (MPhil Stud)

I, Karri Lassi Heikkinen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Roughly, *strong longtermism* (Greaves and MacAskill 2021) is the view that in the most important decision-situations facing us today, the option that is ex ante best, and the one we ought to choose, is the option that makes the far future go best. The purpose of this thesis is to evaluate strong longtermism. I do this by first considering what I take to be three important objections to this view, and then suggesting a way in which the strong longtermist may be able to respond to them.

The thesis consists of five chapters. In Chapter 1, I introduce the topic of the thesis and reconstruct Greaves and MacAskill's argument for strong longtermism. In Chapter 2, I argue that partially aggregative and non-aggregative moral views form a significant objection to Greaves and MacAskill's argument for deontic strong longtermism. In Chapter 3, I discuss the procreative asymmetry, arguing that what I call the Purely Deontic Asymmetry forms another important objection to strong longtermism. In Chapter 4, I consider the problem of fanaticism, arguing that the best way those in favour of strong longtermism can avoid this problem is by adopting a view called tail discounting. Finally, in Chapter 5, I propose that the issues discussed in the preceding chapters can be satisfactorily dealt with by framing strong longtermism as a public philosophy. This means that we should understand strong longtermism as a view that correctly describes what state-level actors ought to do, rather than as a blueprint for individual morality.

If my evaluation is correct, then there are important limits to the role that strong longtermism can play in our private lives. However, it also implies that as a society, we ought to do much more than we currently do to safeguard the long-term future of humanity.

Impact Statement

In intend this work to contribute to the academic literature on our moral duties towards future people, as well as debates between consequentialist and non-consequentialist theories in philosophical ethics more broadly. Beyond academic interest, the questions this thesis addresses are relevant for how we, both as individuals and as a society, ought to conduct our lives. Given that we live in a time where our choices have profound consequences for future generations, such questions are worth thinking about for everyone.

Acknowledgements

I thank my supervisor Joe Horton for his advice, comments and encouragement. I also thank Tim Button, who provided helpful comments on Chapter 4, and my peers at the UCL MPhil Stud Thesis Preparation Seminar, who interrogated me on Chapters 2 and 3. During my time working on this thesis, I have been financially supported by a grant from Jorma and Märtha Sihvola Foundation and a Global Priorities Fellowship from Forethought Foundation, and I am most grateful for them both. I also want to extend special thanks to Ella, Dilara, Liane, Laura and Jelena for being by my side through the ups and downs of this process. Finally, I thank my family—Eija, Jussi and Tuukka—for their unwavering love and support through the years. Without you, I could never have written this thesis.

Table of Contents

1

Introduction

Future people matter. Other beings probably matter too, and future people may not matter exactly as much as those alive now. But the basic moral fact that future people matter, I take it, should be clear to everyone from a myriad of actions we take. We aim to reduce the harmful effects of climate change, even though most of those effects will hit the people who come after us hardest. Some of us worry about the national debt, arguing that it is unfair of us to fund our policies at the expense of future generations. We have international conventions in place that protect culturally valuable sites even at the time of war, and we spend a great deal of resources to preserve works of literature, art and science in museums, libraries and universities. While all these actions can to some extent be justified by other considerations, it seems obvious to me that at least one important reason motivating them is the moral concern for future people.

Another fact about future people, even if not always salient in our minds, is that there could be a very large number of them. Consider the following heuristic used by Our World in Data (Roser, 2022). We are mammals, and it turns out that the typical lifespan of a mammalian species on Earth is about 1 million years. Since humans have existed for about 200,000 years, we might guess that humanity has about 800,000 years left. Now, if the world population stabilises at about 11 billion people, as the UN estimates, the future would include 100 trillion humans—that is one followed by fourteen zeros. Furthermore, note that the Earth might remain habitable for another billion years, which is thousand times the 1-million-year typical mammalian lifespan. So, the above estimate implicitly assumes that we will only manage to survive 0.01% of the time that the Earth remains habitable. And yet, the number of future people would vastly outweigh the number of people alive now.

If future people matter, and if there is a very large number of them in expectation, then it seems like we should take great care to ensure that the future of humanity goes well. Indeed, reading the list of examples above might make it seem like we already do this. But taking a closer look at the world around us tells a different story. Our public discussion follows the 24-hour news cycle, and our day-to-day politics rarely reaches even beyond the next electoral term. As numerous examples from Hurricane Katrina to the Covid-19 pandemic have showed, we are woefully unprepared to deal with risks that materialise in timespans longer than a few decades. And despite some efforts to the contrary, our carbon-intensive way of life risks leaving future generations with a radically impoverished planet.

Many philosophers hold the view that our neglect of the future constitutes a moral error of some kind. In this thesis, however, I focus on just one such view. This view, defended by Greaves and MacAskill (2021), is called *strong longtermism*. In slogan form, strong longtermism amounts to the claim that "impact on the far future is the most important feature of our actions today" (Greaves and MacAskill 2021:1). The purpose of this thesis, as the title suggests, is to evaluate this view. In the remainder of this introductory chapter, I introduce Greaves and MacAskill's position in detail and explain why I take it to be worth exploring. I then offer a roadmap for this thesis.

1.1. Strong Longtermism

1.1.1 Clarifying the view

While the compact statement of strong longtermism above is helpful, it is not yet precise enough to allow rigorous argument. Our first task, therefore, is to dive into the precise definition of strong longtermism that Greaves and MacAskill provide. According to this definition, strong longtermism consists of the following two claims.

> **Axiological strong longtermism (ASL):** In the most important decision situations facing agents today,
>
> i) Every option that is near-best overall is near-best for the far future.
> ii) Every option that is near-best overall delivers much larger benefits in the far future than in the near future.
>
> **Deontic strong longtermism (DSL):** In the most important decision situations facing agents today,
>
> i) One ought to choose an option that is near-best for the far future.
> ii) One ought to choose an option that delivers much larger benefits in the far future than in the near future.
>
> (Greaves and MacAskill 2021:3, 26)

Following Greaves and MacAskill, I will refer to axiological strong longtermism as ASL and deontic strong longtermism as DSL. I will also refer to the conjunction of these claims simply as strong longtermism. Let us begin by unpacking ASL.

ASL is a claim about axiology, meaning it is a claim about value. The first important thing to note is that the claim here is about *ex ante* value, rather than *ex post* value. This means that ASL is concerned with the value of options at the time of decision, based on the beliefs that the agent should have about the decision situation she faces, rather than the objective value of how things ultimately turn out. For most of this thesis, I identify ex ante value simply as expected value. However, I should note that while Greaves and MacAskill also follow this assumption for the most of their paper, the definition of strong longtermism given here is logically compatible with other theories of ex ante value as well.

Throughout the definition of strong longtermism, “far future” refers to the future from some time t onwards which is counterintuitively far away from the time of the decision. Here, time t could be, say, 100 years—Greaves and MacAskill leave it for the reader to decide. Claim i) of ASL then amounts to saying that every option that realises, in expectation, something sufficiently close to the maximum available value overall is such that it still realises, in expectation, something sufficiently close to the maximum available value even if we disregard all its near-term effects. In other words, we can identify the near-best options overall by focusing only on far future effects. Claim ii) then states that this is because the benefits of these overall near-best options are predominantly realised in the far future. In other words, it is *not* the case that the near-best options for the far future are also near-best overall because they somehow happen to be near-best in the short term. Even if some option x is better than another option y in the short term, but option y is better in the far future, then option y will be better overall.

Admittedly, the explanation here is complicated, as is the definition it tries to explain. Luckily, however, we can put the essential part of ASL in simpler terms. As a shorthand, we can say that when it comes to axiological, ex ante value, *the option that makes things go best is the option that makes the far future go best*. While this way of summarising sacrifices some of the preciseness of Greaves and MacAskill's definition, I believe that it helpfully expresses the key thought that Greaves and MacAskill want to argue for.

Having clarified ASL, we can now move on to DSL. Note that ASL is a purely axiological claim, meaning that it does not tell us what we have duties or obligations to do, or what we ought to do all things considered. This 'overall ought' is captured by DSL instead. According to Claim i) of DSL, morality requires that, in the most important decision-situations facing us today, we choose an option that is near-best for the far future. Claim ii) then adds that we ought to choose options that realise, in expectation, much more value in the far future than in the short-term future—again, this rules out the possibility that we ought to choose the ex ante near-best option for the far future just because this also happens to be the near-best option for the short term. Note that while Greaves and MacAskill defend both ASL and DSL, it could be that either one of them is true while the other is false.

Importantly, the ought of DSL is a subjective, rather than an objective ought, similarly to how ASL deals with ex ante rather than ex post value. In other words, we are interested in what an agent should do relative to the beliefs they have (or should rationally have) about the decision they face at the time of making that decision, rather than which choice would ultimately, as a matter of fact, turn out to be the right one. For example, it could be that you subjectively ought to give someone aspirin to relieve their headache, as you reasonably believe that this will help them, even though you objectively ought not to do this, because unbeknownst to everyone this person has a severe aspirin allergy.

As with ASL, I believe that we can provide a helpful shorthand that captures the essential thought behind DSL. We can say that when it comes to what we subjectively ought to do in the most important decision-situations facing us today, *we should choose the option that we expect to make the far future go best*.

There is one more important point to clarify, namely the "most important decision-situations facing agents today" that Greaves and MacAskill repeatedly refer to. According to the authors, the relevant decision-situations include "those of a society deciding how to spend money with no restrictions as to 'cause area', an individual making the analogous decision, and individual career choice" (Greaves and MacAskill, 2021:4). These are taken to be the most important decisions in the sense of having a great potential impact on the wellbeing of both current and future sentient beings—it is not claimed that they are necessarily important in some sense internal to the agent making the decision. In this context, cause area refers to the problem or aim that a given spending or intervention addresses, such as global poverty, homelessness, animal welfare or nuclear non-proliferation. Even though Greaves and MacAskill (2021:30) suggest that strong longtermism could also apply to a wider range of decisions, they only explicitly commit to the definition above. This restriction, I take it, is meant to both make the empirical argument for strong longtermism more plausible and to sidestep the worry that strong longtermism might, if applied to every single action, be a very demanding moral view.

We have now covered all aspects of the definition Greaves and MacAskill give, meaning that we can summarise strong longtermism as a whole. Using a helpful shorthand once more, I believe we can say the following: strong longtermism tells us that when it comes to a society or individual spending money with no limit on cause area and an individual choosing a career, the option that is ex ante best, and the option we ought to choose, is the option that makes the far future go best. The first part (about what is best) forms the axiological claim, whereas the second part (about what we ought to do) forms the deontic claim. Again, while this way of expressing things slightly simplifies Greaves and MacAskill's definition, I believe it should be very useful for a reader who wants to grasp the most essential thought behind strong longtermism.

Having defined strong longtermism and done my best to offer an initial explanation of this definition, I will now move on to introduce Greaves and MacAskill's argument for the view in the following two subsections. After that, I will then use the final section of this chapter to explain my motivation for evaluating strong longtermism and to set out my plan for the rest of this thesis.

1.1.2. Arguing for ASL

The authors begin by arguing for ASL, using a combination of a plausible empirical assumption and a set of ethical claims. Let us begin from the empirical assumption—as it happens, we have already seen it. This is the assumption that the long-run future is vast, and thus involves an overwhelmingly large number of morally significant beings. Here, Greaves and MacAskill (2021:6-9) make the case that even our most conservative estimate should place the number of future people at about 100 trillion (10^{14}), the figure I also mentioned above, while our main estimate should be even higher. The reader is instructed to read their paper for details—for now, what is important is that Greaves and MacAskill take their argument to go through even if we assume the most conservative estimate.

Let us now move on to the ethical claims. Here, we need to be careful: while Greaves and MacAskill put in place some evaluative assumptions to illustrate their view, they explicitly state that some of these assumptions could be replaced by different ones while still getting strong longtermism as a result. The inessential assumptions, as the authors see it, are 1) identifying ex ante value with expected value, in the way that standard expected value theory (EVT) tells us, and 2) identifying value with total welfare, in the way that total utilitarianism tells us.

A brief note on how inessential these assumptions truly are is in order. To begin with, I want to flag that I will discuss the role of EVT in strong longtermism in more detail in Chapter 4 of this thesis. On the other hand, the way in which we can relax the assumption of total utilitarianism should hopefully be relatively clear: we could, for example, assign some value to equality or scientific advances, and note that the overall value of the history of humanity along these dimensions too depends overwhelmingly on how the far future goes. There is, however, one component of total utilitarianism that, for reasons we will soon see, seems to be essential for Greaves and MacAskill's argument. This is the idea of zero rate of pure time preference, which in this context means that, as far as morality is concerned, all consequences of our actions matter equally, regardless of how far away in time these take place. As the authors note, however, this essential assumption seems to be widely accepted among philosophers (see also Greaves 2017).

Keeping all this in mind, let us now put together the argument for ASL. In its simplest form, the idea is this: given that all the consequences of our actions matter equally, and that the size of the long-run future is massive compared to the short term, the amount of ex ante value we can produce by influencing the long term will be larger than the amount of ex ante value we can bring about by influencing the short term. To see the idea, consider an archery competition where you get to shoot a thousand arrows, and each one of them counts equally. It would be very unlikely that you score more points with your first 10 arrows than with the last 990 arrows.

While my reconstruction above captures the basic gist of the argument for strong longtermism, one might think that it is too quick. What if we cannot causally affect the far future? To use the archery analogy again, we might model the idea that it is much harder to influence the far future than the near future by saying that after each arrow you shoot, the target is taken 10 yards further. This could make it the case that you indeed score more points with your first ten arrows than the 990 ones that follow, for after those first ten attempts, it becomes almost impossible to hit the target at all.

Foreseeing a natural objection like this, Greaves and MacAskill argue that when it comes to the very most important decision-situations identified before, there are ways in which we *can* attain far future ex ante benefits that are much higher than the highest attainable near future ex ante benefits. These interventions include work aimed at decreasing the risk of human extinction and scenarios where humanity may end up permanently locked into a bad state of affairs, as well as research into better understanding these risks and building capacity to act on them later. To be more precise, the list here could include things like research and advocacy aimed at making emerging technologies safe; regulation aimed at decreasing the risk of deadly pandemics; developing asteroid shields; stricter climate policy; and perhaps saving resources for future generations to use if they face a time of perils.

In summary then, the argument for ASL can be put as follows. Given that 1) the future is vast, 2) the consequences of our actions matter just as much in the far future, and 3) we can expect to influence the far future for the better, the best thing we can do will be one of the things that most improve the far future. Even if this is unlikely to hold for every decision we ever make, Greaves and MacAskill argue that it does hold for the most important decisions they identify.

1.1.3. Arguing for DSL

When it comes to DSL, Greaves and MacAskill do not argue for this claim directly. Instead, after defending ASL, they give an argument that tries to establish DSL by using ASL as a premise, along with two other premises. To understand the argument below, think of side constraints as things that, according to non-consequentialist views, we are never allowed to do, and personal prerogatives as things that we are always permitted to do. The argument is as follows.

> (P1) When the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor, one ought to choose a near-best option.
>
> (P2) In the most important decision situations facing agents today, the axiological stakes are very high, there are no serious side-constraints, and the personal prerogatives are comparatively minor.
>
> (C) So, in the most important decision situations facing agents today, one ought to choose a near-best option.
> (Greaves and MacAskill, 2021:27)

To complete this, we add ASL as a premise and note that all near-best options are in fact near-best for the far future, giving us DSL. Essentially, the argument above shows that when certain conditions hold, we ought to do what makes things go best in expectation, and the thing that makes things go best in expectation is just whatever makes the far future go best in expectation. Therefore, we ought to do what makes the far future go best in expectation.

Let us look at the premises of the above argument one at a time. For the first premise, Greaves and MacAskill argue that many of the kind of non-consequentialist restrictions or prerogatives, which are usually taken to be the reasons for not choosing the option that is best in the axiological sense, do not retain their power when the stakes are extremely high. This idea—call it the *Stakes Principle*, following Mogensen (2019)—has some initial plausibility. For example, it is plausible to think that while we ought not to lie in normal situations, we should break the rule against lying if the situation is very dire—say, when ten people stand to die unless you tell a lie to a murderer. Furthermore, the premise as expressed here is fairly weak: it still allows the existence of side constraints and prerogatives that are so significant that they trump the obligation to choose a near-best option. For example, we might think that we should not torture innocent children even if this is what makes things go best overall.

For the second premise, consider again the magnitude of the future. If one accepts this empirical assumption, together with Greaves and MacAskill's argument that we can meaningfully influence the vast number of sentient beings it includes, then it seems clear that whether we choose to do what Greaves and MacAskill instruct us to do is a choice with massive axiological stakes. Greaves and MacAskill also believe that when it comes to the kind of decision-situations they single out, there are no side constraints or prerogatives in play that would be serious enough to block the argument. The thought here seems to be that none of the interventions they recommend require us to commit any of the kind of very severe wrongs that some people consider categorically forbidden, such as torturing or killing innocent people.

We can now see the argument for DSL in full. Given the massive axiological stakes of our decision between a near-best option and some other option, most non-consequentialist restrictions against choosing the near-best option have no force. Given also that there are no very serious restrictions that would tell against choosing any of the near-best options in play here, what we overall ought to do is to pick a near-best option. Finally, given ASL, all near-best options overall are near-best for the far future. Thus, we have arrived at DSL. In essence, what happens here is that the Stakes Principle forces a convergence between ASL and DSL by ruling out the kind of factors that in normal situations might separate what we ought to do from doing what makes things go best.

1.2. Evaluating Strong Longtermism

1.2.1. Why strong longtermism?

The purpose of this thesis, as already mentioned, is to evaluate strong longtermism. But why write a whole thesis on what might seem, at least on the surface, to be just a single working paper? Here, three motivating reasons are worth noting.

Firstly, strong longtermism as a philosophical position is deeply rooted in the tradition of consequentialist moral philosophy, meaning that many of the things I say in this thesis can be used to elucidate, defend and critique previous work in this tradition as well. Indeed, the basic idea behind strong longtermism goes back to at least Parfit's (1984) seminal work on population ethics. More recently, versions of the longtermist position have been defended by, among others, Bostrom (2003), Beckstead (2013), Ord (2020) and MacAskill (2022). In responding to this line of work, I hope this thesis can contribute to existing debates between consequentialist and non-consequentialist theories, particularly in the context of future people.

Secondly, I believe strong longtermism is worth engaging with because it goes further than much of the preceding work in the same tradition in many interesting ways. An important part of Greaves and MacAskill's paper is what we might call a 'moral sensitivity analysis': after setting up the basic case for strong longtermism based on total utilitarianism and expected value theory, they discuss many deviations from these assumptions, aiming to argue that many moral views should converge behind strong longtermism. Indeed, Greaves and MacAskill do not assume the truth of any form of maximising consequentialism at the outset. Rather, the Stakes argument is meant to show that even non-consequentialists should accept strong longtermism. This, I believe, is a philosophically ambitious claim that warrants closer inspection.

Thirdly, whether we accept strong longtermism has important practical implications. If strong longtermism is true, then we ought to spend much more than we currently do on reducing existential risks and aiming to make sure the future has a positive value. I already mentioned some of the practical implications above—think AI safety, biosecurity work, asteroid shields and the like. Note that while these interventions may not have any significant short-term effects, there is a reasonable chance that they have major effects in the long run. Therefore, strong longtermism would often recommend these interventions over more traditional ways of doing good in the present, such as donating to the global poor. And finally, beyond such individual choices, we might also think that strong longtermism implies vast social change: perhaps future generations should be represented in our parliaments, their protection written down in our constitutions, and more.

1.2.2. A plan for this thesis

Having introduced the topic of this thesis and explained my motivation, I will now offer a brief description of what is to come. In short, the plan is to consider what I take to be three important problems for strong longtermism, and tentatively suggest a solution to them. The thesis consists of five chapters, the first of which is this introduction.

In Chapter 2, I argue that there is an important form of non-consequentialist thought that the Stakes argument fails to capture, namely partially aggregative and non-aggregative moral views. This means that contra Greaves and MacAskill, not all plausible non-consequentialist views will converge on DSL, and indeed we might be required to reject DSL altogether.

In Chapter 3, I consider how strong longtermism interacts with the procreation asymmetry, arguing that contrary to how things might seem at first, only one particular understanding of the asymmetry is unavoidably in conflict with the view. This understanding, however, seems to me to be the most plausible one, meaning that the asymmetry gives us another reason to be sceptical of strong longtermism.

In Chapter 4, I consider the objection that since strong longtermism relies on assumptions that imply a view called fanaticism, we should reject these assumptions, and thus reject strong longtermism. Here, however, I argue that we can modify the assumptions behind strong longtermism in a way that avoids fanaticism and still allows the strong longtermist argument to go through, by adopting a view called tail discounting.

Finally, in Chapter 5, I propose that while the objections considered in this thesis should be taken seriously, we can respond to them by framing strong longtermism as a *public philosophy*—meaning that we should understand strong longtermism as a view that correctly describes what state-level actors ought to do, rather than a code governing private interactions between individuals.

2

The Challenge from Anti-Aggregative Views

According to Greaves and MacAskill, even most non-consequentialists should accept their view. It is worth noting just how significant a result it would be if their argument turns out to be sound. It is not difficult to argue that the long-run future should be our priority, if we start by assuming utilitarianism (see Bostrom 2003). However, if Greaves and MacAskill are correct, then *any* plausible non-consequentialist theory will also agree with strong longtermism.

In this chapter, I argue that strong longtermism is incompatible with a range of non-aggregative and partially aggregative moral views. The purpose of my argument is to show that, contrary to what Greaves and MacAskill claim, not all plausible non-consequentialist views will converge behind the idea that we ought to do what makes the far future go best in expectation. While side constraints and personal prerogatives—which Greaves and MacAskill explicitly include in their argument for DSL—are a vital part of contemporary non-consequentialist thought, there is also a further important class of considerations that many non-consequentialists care about, namely considerations on how we should aggregate moral claims across individuals.

The chapter has five sections. In Section 1, I briefly explain the variety of views that different authors have taken on interpersonal aggregation. In Section 2, I present a hypothetical case to show that strong longtermism is at odds with non-aggregative and partially aggregative views and reject an initial attempt to dismiss this conflict. In Section 3, I present another case, showing how strong longtermism conflicts with non-aggregative and partially aggregative views in situations involving risk. In Section 4, I locate the source of the issue at the so-called Stakes Principle that Greaves and MacAskill use in their argument. Finally, in Section 5, I address a passage where Greaves and MacAskill seem to suggest that deontic non-aggregative and partially aggregative views are so clearly mistaken that they do not pose a problem for strong longtermism. I argue that Greaves and MacAskill fail to show this.

2.1. Interpersonal aggregation

2.1.1. Introducing aggregation

A good way to get a grip of the different views on interpersonal aggregation is to consider the following. Imagine you can save either one person from a very serious harm x, or n people from some less serious harm y, such that n > 1. Other things equal, is there any number n such that you should save the group of people suffering the less severe harm rather than the one person suffering the more serious harm?

First, the so-called non-aggregative views tell us that the answer is always no (e.g. Taurek 1977). Those who hold this view think it is not appropriate to aggregate the lesser harms of the many to justify saving them over the one person who stands to suffer the most. Instead, what we ought to do in a situation like this is to meet the strongest individual claim, regardless of how large a number n might be. In other words, we ought to satisfy the strongest claim, no matter how large the sum of the competing claims may be.

Second, on the other end of the spectrum, fully aggregative views always give us a positive answer (e.g. Horton 2018). Proponents of fully aggregative views reason in a manner opposite to the non-aggregationists. To find out which group of people to save, we simply add up the (strength-weighted) claims on both sides and see which group comes out on top. It follows (given some other very plausible assumptions) that even if harm x is much more serious than harm y, there always must be *some* number of people suffering y that outweighs one person suffering x.

Finally, one can adopt a partially aggregative view (e.g. Kamm 1993; Scanlon 1998; Voorhoeve 2014; Tadros 2019). The basic idea of all partially aggregative views is that the answer to the above question depends on how similar harms x and y are to each other. In cases where the harms are sufficiently similar, such as when ten people facing severe disability are compared to one person facing death, aggregation is permitted, and we should save the many. In cases where the harms are very far apart in terms of seriousness, such as when one death is compared with any number of people experiencing temporary headaches, aggregation is not allowed, and we should save the person suffering the most severe harm.

It is not possible to evaluate these views in much depth here. For the purposes of this chapter, I want to simply note that both non-aggregative and partially aggregative views have some significant benefits and at least partially aggregative views enjoy a considerable support among philosophers. Therefore, in so far as strong longtermism is at odds with these views, this forms a noteworthy challenge to Greaves and MacAskill's view.

2.1.2. Axiological and deontic versions

Given the distinction between axiological and deontic strong longtermism, it is important to note that this distinction is also present in the aggregation debate. In particular, non-aggregative and partially aggregative views can be interpreted in either deontic or axiological terms. We might think, in a purely axiological sense, that no number of minor harms or benefits can ever add up to enough (dis)value to outweigh the more serious harm or benefit. Alternatively, we could say that while the axiological value of a major claim can be outweighed by a very large number of minor claims, we nevertheless ought to satisfy the major claim, say, because of some deontic requirement of respect towards each individual.

From now on, I will use the term *anti-aggregative* to denote the set of all non-aggregative and partially aggregative views. This is meant to improve readability and allow us to easily distinguish between the set of all anti-aggregative views and its subsets.

2.2. Lisa's Choice

2.2.1. The argument

We are now in a place to see the conflict between strong longtermism and anti-aggregative moral views. To get an idea of the general form of my argument, consider the following case.

> Lisa needs to allocate 1 million dollars. This money came from a will, which stipulates that the purpose of the donation is to provide medical treatments in the US. Lisa figures out that she can either provide a lifesaving treatment to one person or use the money to set up a foundation which provides treatments for minor ailments. She knows that with a smart mix of investment and spending, this foundation could go on to operate for hundreds of years and provide treatments to at least a million people. Thus, Lisa faces a choice between saving one person from death now and saving at least a million people from a minor ailment in the long run.

What should Lisa do?

The views considered so far give the following judgements. According to any anti-aggregative view, Lisa should save the one person from death, because the claim to be saved from death takes priority over any number of minor ailments. Fully aggregative views would instead favour setting up the foundation, since the aggregate of a million minor ailments is (by stipulation) enough to outweigh one death.[1]

[1] I assume throughout this thesis that the aggregative views discussed are temporally neutral.

How about strong longtermism? Here, it is important to note that the way I have set up the case means that it falls outside the set of the decision-situations that Greaves and MacAskill concentrate on. However, Greaves and MacAskill contend that strong longtermism might apply to many other decisions too, and this seems like a significant one for Lisa. So, grant for the sake of the argument that Lisa's choice falls under the scope of the view. What can we learn?

I think it is clear that if Lisa believes in strong longtermism, she will choose to set up the foundation. Firstly, if we were not allowed to aggregate harms and benefits across all the future people in a fully aggregative manner, then the fact that there are so many of them would not have the moral significance that strong longtermism takes as its starting point. Secondly, recall that strong longtermism relies on the idea that all effects of our actions count equally. Typically, the way anti-aggregative views handle these kinds of cases is by maintaining that (at least in some contexts) claims that are small enough do not matter at all. In disagreeing with these views, strong longtermism aligns naturally with fully aggregative views.

At this point, the reader might wonder whether any of this makes a difference, given that Lisa's choice still does not fall under the scope of strong longtermism. However, the point of my example is more general: it is meant to show—in the easiest way I can think of—that strong longtermism relies on accepting a fully aggregative moral view. This, in turn, means that we can construct a variety of examples where strong longtermism conflicts with anti-aggregative moral views. Lisa's choice demonstrates a situation where we must choose between (1) delivering a very significant benefit to a small number of people in the present, and (2) setting up some mechanism which will reliably deliver smaller benefits to a very large number of people in the long run. We can imagine all sorts of candidates for what this reliable mechanism could be. For example, following some longtermist effective altruists, we might aim to promote broadly positive values or improve people's reasoning skills, either through targeted education or the development of cognitive enhancements.

2.2.2. Objection: is this relevant?

Even if we accept that Lisa's choice tells us something important about the way in which strong longtermism conflicts with anti-aggregative moral views, there is another natural way in which one might be inclined to dismiss the issue raised by the example. One might think as follows: the most plausible anti-aggregative views, such as that of Voorhoeve (2014) take a deontic form, meaning that they do not pose a challenge to ASL. Furthermore, if ASL is true, then regardless of whether DSL is true, making the long-run future go well is the morally best thing one can do. And for an effective altruist, that is an important result, even if we are not strictly obliged to follow it. Even Mogensen (2019), who argues against DSL, still grants something like this line of thought.

Unfortunately, I believe that the anti-aggregationist critique of strong longtermism is potentially much more damaging. This is because deontic anti-aggregative views—the ones the response assumes to be most plausible—require us to meet the strongest individual claim, regardless of what is best in the axiological sense. Therefore, in cases where strong longtermism and deontic anti-aggregationist views disagree, choosing the longtermist option could turn out to be impermissible. In such a situation, ASL would have no action-guiding significance, even if it was in some sense true. Thus, anti-aggregationism poses a serious challenge to both the truth of DSL and the action-guiding power of ASL. Furthermore, it also calls into doubt the soundness of the argument from ASL to DSL.

2.3. Sarah's Choice

2.3.1. The case

Lisa's choice illustrates what I believe to be the simplest way in which anti-aggregationism and strong longtermism can come apart. However, I believe that there is also another, much more worrying type of situation where the two views go against each other. Consider the following:

> Sarah wants to donate 1 million dollars to improve the world as much as she can. She has narrowed the choice down to two options: either donate to GiveDirectly or the Machine Intelligence Research Institute. Donation to GiveDirectly would make a direct positive contribution to the lives of about a thousand households living now. Donating to MIRI would amount to betting on a very small probability of significantly improving the lives of an extremely large number of future people. Thus, Sarah faces a choice between giving a certain, significant benefit to a relatively small number of people living now and giving a very small probability of a significant benefit to an extremely large number of people in the long run.

For context, GiveDirectly provides unconditional cash transfers to extremely poor families in Uganda, whereas MIRI does fundamental research on AI. In the effective altruist community, the latter would be considered a paradigm case of a longtermist intervention, whereas the former is seen as a way to help people in the present. What should Sarah do?

On this question, strong longtermism sides with fully aggregative views, for the same reasons as before. Assuming that donating to MIRI brings about higher expected aggregate value over the course of the long-run future, strong longtermism must prefer this option over donating to GiveDirectly. But from the point of view of anti-aggregative views, the question is slightly more complicated, as I will now explain.

2.3.2. Anti-aggregative views and risk

From an anti-aggregative point of view, it is not immediately obvious whether we should evaluate moral claims ex ante or ex post in situations involving risk. However, I believe that Frick (2015) offers a convincing argument in favour of the former option. We can present a simplified version of his main argument in the form of the following case:

> We need to vaccinate 1 million children to stop them all from dying. Vaccination Program A is 99% effective against the disease, meaning that (roughly) 10,000 children will not survive the pandemic, but we have no idea who these children will be. Vaccination Program B is 100% effective against the disease, but it also has lethal side effects to children who carry a certain, well-known gene. As it happens, we know that 10,000 children in our cohort carry this gene, and we know who these children are. However, unless we give everyone either vaccine A or vaccine B, everyone will die. Which program should we choose?

The problem with ex post anti-aggregative views is that they cannot distinguish between the two programs. This is because whichever vaccine we choose, (roughly) 10,000 children will die, meaning that on the ex post reading, the options are in the morally relevant way identical. The ex ante reading, on the other hand, provides the intuitively correct answer that we ought to choose program A. This is because under program A, every child faces a 1% risk of death, whereas under program B, some children face certain death. Therefore, these children have a weighty moral claim against program B, while no one has a similar claim against program A.[2]

Adopting the ex ante reading means that we face situations where a certain benefit (or harm) must be compared with an uncertain one. If the

[2] For simplicity, I assume here that there is no way to acquire any further information about the effects of the vaccine programs or the identities of the children who stand to die. However, I should note that relaxing this assumption may cause problems for the ex ante reading. Frick (2015) discusses such cases in more depth.

benefits in question are equal in magnitude, then the certain benefit generates a weightier moral claim than the uncertain one. It follows that, according to anti-aggregative views, we should sometimes give a small number of people a certain benefit rather than giving a large number of people a tiny chance of receiving a benefit of similar magnitude.[3] This is because providing the certain benefit meets the ex ante weightiest moral claim.

The situation described above is exactly the situation that my example involving Sarah tries to capture. Both receiving a cash transfer in the context of extreme poverty and living in a world with safe artificial intelligence are supposed to be major benefits, but the former applies to a relatively small number of people with certainty, whereas the latter would involve a vast number of people with a tiny chance of receiving the benefit. In this case, I hold that, contrary to strong longtermism, the most plausible anti-aggregative views would tell Sarah to donate to GiveDirectly.

Interestingly, the way anti-aggregative views can give some degree of extra weight to certain, identifiable benefits over statistical ones[4] might allow us to explain the intuitive discontent that one might feel about the idea that doing the most good requires taking very tiny chances of landing very valuable outcomes. Effective altruists have tended to consider this to be a general problem in decision theory. However, taking seriously deontic anti-aggregative

[3] Of course, we cannot accept a lexical priority of identified benefits over statistical ones, as this would imply that we ought to sacrifice any number of statistical lives to save one identified victim. Frick (2015: 219-221) is aware of this and argues that we should accept a pluralist account, where the strongest individual moral claim is only a part of what makes acts overall right or wrong. To take this aspect into account in our present example, I should stipulate that the expected value of donating to MIRI over the long term is higher than that of donating to GiveDirectly, but not by an overwhelmingly large amount. Similarly, in the vaccination case above, if program A was much less effective than program B, then we may be required to choose B even though it leads to 10,000 identifiable children dying with certainty.

[4] By statistical benefits I mean the benefits we can predictably generate by applying some policy over many people, but without us being able to know which of these people will end up receiving the benefits. The idea comes from Schelling (1968).

views implies that at least a part of the problem might fall under moral philosophy instead.

2.4. Diagnosis

2.4.1. The source of the problem

We have now seen both Greaves and MacAskill's argument for strong longtermism and how this stance is in tension with anti-aggregative views. What should we make of this conflict?

I believe the source of the issue can be located at the Stakes Principle. The Stakes Principle maintains that if the axiological stakes are high enough, deontic limitations tend to subside. However, for many philosophers, the relationship between axiology and what we ought to do is more complicated. In particular, from the point of view of anti-aggregative views, how the moral stakes in question are distributed among the people involved is of crucial importance.

Some examples will be helpful here. Note that the kind of cases that make the Stakes Principle seem plausible have the feature that the individual claims on each side are either equal or weigh in favour of breaking the relevant non-consequentialist constraint. Consider, for example, lying to save the lives of ten people, or killing one person to save thousands. In the first case, being saved from death is much stronger a claim than the claim to not be lied to. In the second case, each individual claim is identical. In cases like this, the Stakes Principle is not in conflict with (most) anti-aggregative views, and we are left with little reason for scepticism.

Now, compare these cases to another imagined case where we can make sure no one in the history of the universe will ever have a broken nail, but this comes at the expense of leaving one person to die. The difference here should be obvious: the Stakes Principle is now at odds with anti-aggregative moral views. The Stakes Principle assumes that the overwhelmingly large stakes overriding side constraints and prerogatives can be made out of an arbitrarily large number of arbitrarily small individual claims. But this is precisely what all those in favour of anti-aggregative views deny, even if they do not deny axiology altogether.[5]

2.4.2. What next?

Faced with this challenge, the longtermist can respond in two ways. The first option is to limit the scope of strong longtermism to cases where there are no anti-aggregationist reasons to prefer a short-term intervention. Perhaps we can add a condition to DL, which stipulates that the moral claims of the future people that we expect to meet have to be sufficiently similar to the strongest claims of people living among us now. Essentially, this would amount to integrating a deontic partially aggregative component into strong longtermism. The wider implication of this thought might be that we should pivot from unqualified longtermism towards caring more about so-called *suffering risk*—that is, aiming to make sure that the very worst kind of future scenarios never take place (Bauman 2017).

The second option is to argue that both axiological and deontic anti-aggregative views are implausible. This response sacrifices some of Greaves and MacAskill's ecumenical ambitions, but if successful, it could strengthen the case for strong longtermism. While many philosophers will resist this move, some recent work suggests that fully aggregative views have important benefits over anti-aggregative views (e.g. Tomlin 2017; Horton 2020). Indeed, in my opinion, this is ultimately the preferrable route.

My reasoning behind this view is the following: if the longtermist tries to accommodate anti-aggregative views, then it will be difficult to find real-life

[5] Some anti-aggregative views do deny axiology altogether. But as I show here, we need not go that far to get into problems.

interventions that would be recommended by the revised view. It seems to me that almost all real-life longtermist interventions suggested so far involve either (1) setting up a mechanism that will reliably deliver relatively small benefits over the long term, as in Lisa's choice, or (2) giving a tiny chance of a major benefit to a very large number of people in the long term over giving a certain major benefit to a smaller number of people in the present, like in Sarah's choice.

One case that I think illustrates this dynamic particularly starkly is when longtermist interventions are aimed at reducing the risk of human extinction. Firstly, it is plausible that to the individual herself, being brought into existence is a benefit, but not as great a benefit as, for example, being freed from extreme poverty.[6] Prioritising extinction risk reduction over the elimination of global poverty may, therefore, exhibit the same old structure of giving a smaller benefit to a very large number of people over giving a greater benefit to fewer people. In this sense, the very survival of humanity may be analogous to Lisa's foundation: it is a mechanism which will reliably produce a very large number of moderate benefits over the long term.

Note that the issue here is not about person-affecting views in population ethics, which may recognise no reason for avoiding human extinction at all. Rather, my point is simply that in deciding which moral claims to prioritise, an anti-aggregationist agent may well view extinction prevention as much less important than strong longtermism implies, because there are more weighty individual claims than the claim to be brought into existence.

Secondly, there is the obvious fact that any intervention aimed at reducing existential risk involves betting on small probabilities of success. Therefore, existential risk as a cause area involves both of the two mechanisms illustrated by Lisa's and Sarah's choices, where anti-aggregationist thinking goes against

[6] Some readers will want to deny the idea that coming into existence is a benefit. But denying this will not help the strong longtermist, and the point here is to articulate a problem they face even with favourable assumptions.

strong longtermism. However, this cause area also seems so central to the strong longtermist project that letting it go is too big a price to pay for ecumenicism. The only tenable option for proponents of strong longtermism, then, is to argue against anti-aggregative views.

2.5. Are anti-aggregative views obviously implausible?

2.5.1. Greaves and MacAskill's objection

There is one more objection I should address. It seems to me that part of the reason why Greaves and MacAskill devote so little attention to anti-aggregative views, despite their overall aim of making strong longtermism compatible with a wide range of ethical assumptions, is that they think anti-aggregative views are simply mistaken. If this was true, then those in favour of strong longtermism could safely ignore anti-aggregative views. I am inclined to agree with this assessment when it comes to axiological anti-aggregative views. However, as I will now explain, I do not think that Greaves and MacAskill succeed in showing that the same holds for deontic anti-aggregative views.

When discussing their argument for DSL, the authors make the following brief remark concerning a possible objection against it:

> Let's first consider the non-aggregationist response. Consider a Briton, during WWII, deciding whether to fight; or someone debating whether to vote in their country's general election; or someone deciding whether to join an important political protest; or someone deciding whether to reduce their carbon footprint. In each case, the ex ante benefits to any particular other person are tiny. But in at least some such cases, it's clear that the agent is required to undertake the relevant action, and the most natural explanation of why is because the axiological stakes are so high. (Greaves and MacAskill, 2021:28)

The thought here seems to be that since non-aggregative (or anti-aggregative, as I call them) views entail that small benefits are not relevant in determining what we should do, these views imply that we have no reason to do any of the things Greaves and MacAskill mention. But clearly, we do have such reasons, so anti-aggregative views must be mistaken.

This argument does not strike me as particularly persuasive. The core idea of the most plausible deontic anti-aggregative views, such as the partially aggregative view defended by Voorhoeve (2014), is not that small benefits do not matter, but rather that whether they matter in a given context depends on how they compare to other relevant moral claims. For example, these views imply not that you never have a reason to vote, but rather that, if voting somehow came at the cost of leaving a person to die, then you ought to save this person instead. To put it in another way, saying that you should not cure a thousand headaches at the expense of leaving one person to die is not to say that headaches never matter at all. The claim is just that when you are forced to choose, saving a person from death takes precedence.

It seems that there is also a terminological issue here, which may cause confusion. Throughout their paper, Greaves and MacAskill talk about *non-aggregative* views, even though their citations on page 28 show that what they mean is both non-aggregative and *partially aggregative* views. In order to disambiguate this, I chose the term *anti-aggregative* to denote the set of all non-aggregative and partially aggregative views. What makes things messy is that the argument Greaves and MacAskill give may be a good move to make against some very strict non-aggregative views, understood as a subset of anti-aggregative views. It is not, however, a very good argument against partially aggregative views, and consequently, it cannot refute all anti-aggregative views in general.

2.5.2. A different problem?

In fact, it seems to me that Greaves and MacAskill are raising a whole different moral problem in the passage just quoted—namely, the one known as the problem of collective harm, or the inefficacy problem. Roughly, the problem is that there are situations where by acting in some way, we collectively cause great harm (or fail to prevent it), but no individual act seems to make any difference, so it is very difficult to explain why any of these acts would be wrong. Oft-cited examples of situations like this include some of the examples Greaves and MacAskill mention, such as voting in elections or reducing one's carbon footprint to mitigate climate change (for an overview, see Nefsky 2019).

While this is a difficult moral problem, I cannot see how bringing it up would lead us to the conclusion that all anti-aggregative views are implausible. Firstly, fully aggregative views must deal with the very same problem. One way in which those in favour of fully aggregative views could try to solve the inefficacy problem would be to insist that we can aggregate the effects of many agents' actions. But such an idea is not truly central to fully aggregative views—the interpersonal aggregation debate is typically understood to concern aggregation at the level of patients, rather than actors. Consider, for example, the stylised cases where we can save either one person from death or ten people from losing a limb.

Secondly, in so far as deontic anti-aggregative moral views naturally coincide with non-consequentialism, many people would think that these views can deal with the inefficacy problem in ways that are not available to their opponents. Perhaps, for example, voting is simply your moral duty as a citizen, regardless of how large or small the resulting benefits might be. Personally, I would not want to commit myself to this view, but it serves the purpose of demonstrating that, as things stands, Greaves and MacAskill do very little to show the untenability of anti-aggregative views. Thus, anti-aggregative views remain relevant. They are not so obviously mistaken that those in favour of strong longtermism could simply ignore them.

Conclusion

In this chapter, I have shown that strong longtermism conflicts with anti-aggregative moral views. The most important implication of this result is that contrary to what Greaves and MacAskill argue, we should not expect all, or even most non-consequentialist views to converge behind strong longtermism. Strong longtermism relies on fully aggregative reasoning, and as such, it is incompatible with anti-aggregative views. Furthermore, in so far as the most plausible anti-aggregative views take a deontic form, pointing out that these views might not threaten ASL has a very limited relevance with regards to how we ought to spend our resources.

3

The Challenge from Procreative Asymmetry

It seems very likely that realising anything close to a near-best scenario for the far future must inevitably involve creating new people. After all, we are far away from being able to stop aging, and it would be very surprising if a world without humanity would carry more axiological value than a world where humanity flourishes. So, strong longtermism seems to imply that we ought to create new happy people—at least, bearing in mind the range of decisions that the view is meant to apply to, in the sense that a society should spend resources on making sure new people are created. However, this implication seems worrying to many. It is very intuitive to think that even if we have an obligation to make existing people happy, we have no obligation at all to create new happy people. This intuition is known as the procreative asymmetry, or following McMahan (1981), simply the Asymmetry.

In this chapter, I explore how the Asymmetry may be used to argue against strong longtermism. The chapter has three sections. In Section 1, I introduce the Asymmetry and distinguish between what I take to be the two most plausible versions of the view. In Section 2, I investigate which of these versions, if any, is incompatible with strong longtermism. In Section 3, I ask whether we have good reasons to take the Asymmetry seriously, arguing that when it comes to the version of this view that is most problematic for strong longtermism, the answer is yes.

3.1. The Asymmetry

3.1.1. Introducing the Asymmetry

In order to get a grip of the Asymmetry, we can begin with Roberts' statement of it:

> *Claim 1.* It would be wrong to bring a miserable child—a child whose life is less than worth living—into existence.
>
> And:
>
> *Claim 2.* It would be permissible not to bring a happy child—a child whose life is worth living or even well worth living—into existence. (Roberts 2011:765)

Both these two claims are highly intuitive. To begin with, we act in many ways that reflect Claim 1. For example, we use a great deal of resources to monitor the health of a foetus and the mother carrying the child, aiming to make sure that the child will be born healthy. We also morally condemn those who, either through their own fault or serious negligence, risk bringing into existence a miserable child. On the other hand, the idea that we would be obligated to procreate, even if we knew our child would live a happy life, strikes most of us as absurd. This is exactly what Claim 2 says.

What makes these two intuitions asymmetrical in a puzzling way is that they seem to reflect different kinds of moral attitudes. We can see the tension by asking ourselves the following question: if the misery of a potential future child generates an obligation to not create this child, why does another potential child's happiness *not* generate an obligation to bring her into existence? In ordinary situations, both misery and happiness have great moral significance. But when it comes to procreation, we think that while it is wrong to create miserable lives, it is not wrong to not create happy lives.

To be clear, many philosophers reject the Asymmetry (e.g. McMahan 1981, Singer 2011, Parfit 2011). Furthermore, even those who do find it compelling admit that the Asymmetry calls for an explanation of some kind. In this chapter, however, I concentrate on the implications that accepting the Asymmetry has for strong longtermism. In other words, I assume that some form of the Asymmetry is true for the sake of the argument—until Section 3, where I briefly consider whether we ought to accept the version of the Asymmetry I believe is most problematic for those in favour of strong longtermism. Given just how intuitively appealing the basic idea of the Asymmetry is, as well as the number of philosophers defending some version of it (e.g. Narveson 1973, Roberts 2011b, Frick 2020), it seems to me that the view poses a noteworthy challenge.

3.1.2. Clarifying the asymmetry: promoting happiness

The first clarification that we should make concerns the role that promoting happiness (or wellbeing) plays in the Asymmetry. One natural thought, which fits together with the Asymmetry as stated above, but is not in fact what the Asymmetry is about, goes like this: avoiding misery has a special, obligation-generating moral significance that promoting happiness lacks. If this idea was correct, then it would give an easy explanation for the Asymmetry. Creating miserable lives is wrong simply because it would break our obligation to avoid causing misery. But because we have no corresponding obligation to promote happiness, there is nothing wrong with not creating happy lives.

This conclusion is too quick. We often do have obligations to promote other people's happiness, especially in cases where we can do so at very little cost to ourselves. Those who defend the Asymmetry do not deny this. Rather, they hold that the case of procreation is special in a way which creates asymmetrical obligations. In order to take into account that promoting happiness does have moral significance, we can expand the Asymmetry with the following claim:

Claim 3. We ought to promote the happiness of at least the people who already exist.

Adding this claim makes it clear that the reason that it is permissible to not create happy lives is not because we never have obligations to promote happiness. While we are not obligated to bring into existence happy children, we have strong reasons to care about the happiness of at least those who already exist. Often, those defending the Asymmetry also want to include other people, such as the set of those who will exist regardless of the individual procreative decision in question. However, it turns out to be difficult to specify exactly what this further set of people should be, and in any case, I do not think this question is very important for our purposes here.[7]

3.1.3. Clarifying the asymmetry: moral reasons

The next clarification concerns the deontic status that the Asymmetry assigns to different acts. So far, I have used (and cited) expressions like permissible, wrong, ought to, and obligated. This language is helpful in so far as it connects the Asymmetry with familiar intuitions, such as the idea that we are never obligated to procreate. But strictly speaking, these phrases are also misleading. This is because they make it difficult to set aside certain moral concerns which relate to procreation but are not really what the Asymmetry is about. For example, consider again the claim that it is always permissible to not procreate.

[7] Usually, those defending the Asymmetry do not want to limit their concern solely to already-existing people, but instead, include future people whose existence is either already necessitated in some way, or who will exist independently of the procreative choice in question. However, demarcating exactly the set of people whose happiness matters is a complicated problem, so I will here formulate the Asymmetry by focussing on what all its plausible versions would agree on, namely that already-existing people’s happiness matters. I believe this is harmless, because the exact answer to this problem does not change my main arguments. To see this, we can consider a longtermist world government trying to decide whether humanity, collectively, should procreate at all, or whether the resources used on bearing children should instead be used for something else. In such a global, collective decision, there are no future people who will exist independently of what choice is made.

One natural argument for this claim is that procreation involves people's bodies, and since people have absolute autonomy over their own bodies, it can never be obligatory to procreate.

This argument, even if sound, is not the point of the Asymmetry. Those who defend the Asymmetry would hold onto their claims even if one day technological progress allowed us to create new lives with a push of a button, in some safe and effective facility which did not involve the parents' bodies in any way. To avoid issues like this, it is better to talk about *moral reasons*, rather than permissibility, obligations, or the like. We can reformulate the claims that make up the Asymmetry as follows:

> *Claim 1.* We have strong moral reason to not bring a miserable child—a child whose life is less than worth living—into existence.
>
> *Claim 2.* We have no moral reason to bring a happy child—a child whose life is worth living or even well worth living—into existence.
>
> *Claim 3.* We have strong moral reason to promote the happiness of at least the people who already exist.

Formulating the Asymmetry like this makes it clear that the reason why we normally have no obligation to procreate is not because of some factors that override the happiness of the potential child, but rather because the act of creating a new happy life simply is not the kind of thing that we have a moral reason to do. At the same time, this framing allows there to be other, external moral reasons that may tip the balance the other way.

3.1.4. Clarifying the asymmetry: axiological and deontic asymmetry

The final clarification concerns the difference between what I will call the Axiological Asymmetry and the Deontic Asymmetry. The claims about moral reasons above represent the Deontic Asymmetry. We can condense them into the following statement.

> Deontic Asymmetry (DA): We have strong moral reasons to not create miserable lives and to improve the lives of at least those who already exist, but no moral reason to create new happy lives.

Axiological Asymmetry, on the other hand, is a claim about value, rather than what we have reason to do. We can formulate the claim as follows.

> Axiological Asymmetry (AA): Creating a miserable life makes the world worse and at least improving an already-existing life makes it better, whereas creating a new happy life does not influence the axiological value of the world in any way.

The main point of the Axiological Asymmetry is that creating a happy life does not make the world better or worse. Here, the asymmetry between creating a miserable life and creating a happy life is particularly stark: only the former can affect the amount of value in the world.

Once we distinguish between the deontic and the axiological versions of the Asymmetry, we see that it is also possible to combine these versions in multiple ways. Assuming that both axiological and deontic considerations are sensible (contra Thomson 1997), there are three ways to understand the Asymmetry: either DA and AA are both true, DA is true but AA false, or DA is false but AA true. Let us consider what each of these possibilities amount to.

First, consider the possibility that both the Deontic Asymmetry and the Axiological Asymmetry are true. This seems like a natural alignment between the deontic and the axiological: given that creating new happy people does not make the world better, we have no moral reason to do so. We can call this view the *Full Asymmetry* and define it as follows:

> Full Asymmetry (FA): both Deontic Asymmetry and Axiological Asymmetry are true.

This means that each of the conjuncts that make up the Deontic Asymmetry as well as those that make up the Axiological Asymmetry are all true.

Consider next the possibility that the Deontic Asymmetry is true, but the Axiological Asymmetry is false. Since the Axiological Asymmetry is a conjunction of three claims, it would be false if any one of these claims was false. That said, it is difficult to argue that bringing into existence miserable people would not make the world worse, or that helping already-existing people would not make the world better. So, the thought must be that creating new happy people *does* make the world better. Combine this with the Deontic Asymmetry, and the view we get is that even though creating happy people makes the world better, we nevertheless have no moral reason to do so. We can call this the *Purely Deontic Asymmetry* and define it as follows.

> Purely Deontic Asymmetry (PDA): Creating a miserable life makes the world worse and at least improving an already-existing life or creating a new happy life makes it better. However, while we have strong moral reasons to not create miserable lives and to improve the lives of those who already exist, we have no moral reason to create new happy lives.

In what follows, I will concentrate only on the Full Asymmetry and the Purely Deontic Asymmetry, as these are, in my view, the two most plausible ways of understanding the Asymmetry.

There is, of course, a third logically possible combination, namely the Deontic Asymmetry being false but the Axiological Asymmetry being true. The most natural way that the Deontic Asymmetry could be false is if we in fact *do* have moral reasons to create new happy people. Combining this with the Axiological Asymmetry would amount to claiming that even though creating new happy people does not make the world better, we still have a moral reason to do so. We could call the resulting view the *Purely Axiological Asymmetry* and define it in the same style as the views defined above. However, it strikes me that this view can hardly be called a version of the Asymmetry anymore, namely because it implies that, at least when no competing moral reasons are in play, we ought to create new happy people.

3.2. The Asymmetry and Strong Longtermism

3.2.1. Full Asymmetry

Having clarified the Asymmetry and distinguished between two versions of the view, our next task is to investigate how these interact with strong longtermism. If the Asymmetry is inconsistent with strong longtermism, then our intuitions that support the Asymmetry may give us a reason to reject strong longtermism. Let us begin by investigating whether strong longtermism can be true if the Full Asymmetry is true.

We can begin from the axiological side of things. The Axiological Asymmetry is not incompatible with axiological strong longtermism (ASL). The Axiological Asymmetry simply names one thing that makes the world worse, one that makes the world better, and one that does not affect the value of the world. ASL essentially tells us that, in expectation, whichever option realises the most value over the very long-term future is also the option which realises the most value overall. Under this picture, the Axiological Asymmetry is part of what defines which option makes the long-run future go best, whereas ASL tells us that this option is also best overall. To put it in another way: the Axiological Asymmetry takes an object (a new miserable child, for example) and assigns axiological value to it, whereas ASL begins from something that already has an axiological value (what makes the far future go best) and makes another axiological statement about it. Once we see the different levels that these two claims operate on, it becomes clear that they are not incompatible.

That said, accepting the Axiological Asymmetry would still have interesting implications for real-life longtermist projects. Most importantly, it implies that reducing extinction risk (that is, making sure the humankind will not go extinct) is not as important as many effective altruists currently believe. This is because, according to Axiological Asymmetry, adding new happy people into the world does not make the world better. Instead, the Axiological Asymmetry implies that the most valuable thing longtermists can do is to try avoiding scenarios where very many miserable people are being born with no end in sight. For example, we could imagine a perpetual global dictatorship, where some tyrannical leader powered by advanced military technology enslaves most of humanity.

The Deontic Asymmetry is also compatible with ASL. This should be clear from the fact that the former is a deontic claim, whereas ASL is an axiological claim. Furthermore, the Deontic Asymmetry is clearly compatible with the idea that we should try to avoid a future where countless miserable lives are created, since it tells us that we have a strong reason not to create such lives. Similar remarks apply to the link between the Axiological Asymmetry and DSL. But is the Deontic Asymmetry compatible with Deontic Strong Longtermism (DSL)?

Translating DSL into a claim about moral reasons, we can put it as follows: in the most important decision situations, we have a strong moral reason to choose a near-best option for the far future. Since we are assuming the truth of the Axiological Asymmetry at this point, we know that to make the long-run future go best, we should avoid creating miserable lives and improve the lives of those who already exist, whereas creating new happy people does not matter. Therefore, according to DSL, we have strong moral reasons to not create miserable people, and to improve already-existing lives, but no moral reason to create happy lives. This is exactly what Deontic Asymmetry claims. So, in sum, Deontic Asymmetry and DSL are compatible*, given the truth of the Axiological Asymmetry*.

The upshot of this discussion is that what I named the Full Asymmetry would not force us to reject strong longtermism. However, this would mean that instead of working on reducing the risk of human extinction, we should be working to make sure that the worst imaginable future scenarios never take place.[8]

3.2.2. Purely Deontic Asymmetry

We can now move on to consider the second possibility, namely the Purely Deontic Asymmetry. In short, this is the view that while creating new happy people makes the world better, we nevertheless have no moral reason to create new happy people.

As before, the axiological component of this view does not pose a problem for ASL. The Axiological Asymmetry being assumed false simply means that creating happy people is now a determinant of what makes things go best, for it makes the world better. This claim does not directly threaten DSL either, since the former is an axiological claim, while the latter is a deontic one.

However, the truth of the Deontic Asymmetry now poses a serious problem for DSL. To see this, note that if Axiological Asymmetry is false, then part of what makes the future go best is that we create very many happy people over the long-run future. It follows that according to DSL, we have strong moral reason to make this happen, for example by trying to reduce extinction risks. But according to Deontic Asymmetry, we have no such reason. This means that DSL is incompatible with the Deontic Asymmetry, *given the falsity of the Axiological Asymmetry*. In other words, strong longtermism is incompatible with the Purely Deontic Asymmetry.

This conflict arises from the disconnect between the axiological and the deontic posited by the Purely Deontic Asymmetry: under this view, we have no moral reason to do something that makes the world better, namely create new

[8] Greaves and MacAskill (2021:18) make a similar point in relation to person-affecting views in population ethics in general, suggesting that even under such views, longtermist work aimed at ensuring the safety of artificial superintelligence could still be very valuable. Such work, in their view, amounts to making sure that a very bad future never materialises.

happy people. This does not fit together with strong longtermism, because strong longtermism claims a different, more straightforward link between the axiological and the deontic. According to strong longtermism, at least when it comes to the most important decision situations, we have strong moral reason to choose an option that is near-best overall—this being whatever is near-best for the long-run future. Compare this situation to the one posited by the Full Asymmetry. Under the Full Asymmetry, both the Axiological Asymmetry and the Deontic Asymmetry are taken to be true, meaning that no disconnect arises between the axiological and the deontic.

3.2.3. A further complication and intermediate conclusions

At this point, I should also flag one further complication that I have so far omitted. In the above, I have assumed the Axiological Asymmetry is straightforwardly either true or false, but I have not considered the view that the whole idea of axiology is unintelligible. But it is possible to think that there is no such thing as good *simpliciter*, as, for example, Thomson (1997) famously claims. What should we make of the interaction between strong longtermism and the Asymmetry, if we take the Asymmetry to consist exclusively of the Deontic Asymmetry?

Strong longtermism requires axiology to get off the ground. Therefore, whether DSL is incompatible with the Deontic Asymmetry depends on what axiological theory the strong longtermist adopts. In adopting that theory, we reintroduce axiology into the picture, and the situation becomes the same as before: if, *according to strong longtermism*, creating new happy people makes the world better, then the Deontic Asymmetry conflicts with strong longtermism. Similarly, if, according to strong longtermism, creating happy people does not make the world better, then these views converge.

Of course, this conflict or convergence only takes place on the level of practical implications: either way, those who do not believe in axiology will still think that the strong longtermist is making a mistake on the way, even if they arrive at the same destination. Furthermore, the rejection of axiology might offer a deeper justification for why one accepts the Deontic Asymmetry in the first place, meaning that we should not say that the status of axiology does not matter at all. That said, when it comes to practical disagreement on what we have moral reason to do, it seems to me that views that reject axiology altogether ultimately coincide with either views that accept or reject the Axiological Asymmetry. For this reason, I believe that what I have already said in Sections 2.1. and 2.2. is sufficient for our purposes.

We have now clarified the Asymmetry and the way the different versions of this view can interact with strong longtermism. We saw that whether the Asymmetry gives one a reason to reject strong longtermism depends on how exactly the Asymmetry is understood. I argued that, perhaps somewhat surprisingly, what I called the Full Asymmetry does not force one to reject strong longtermism. However, accepting this view does imply that instead of aiming to make sure there are many happy people in the future, we should make sure that the very worst futures never take place. On the other hand, I also argued that what I called the Purely Deontic Asymmetry is incompatible with strong longtermism, meaning that those who find the former compelling have a reason to be sceptical of the latter.

3.3. Should the strong longtermist be worried about the Asymmetry?

3.3.1. Is Axiological Asymmetry credible?

So far in this paper, I have focused on clarifying how the Asymmetry interacts with strong longtermism. However, I have said little in terms of evaluation. In this section, I ask what the implications of my findings are for strong longtermism.

Before going further, recall the difference between the Full Asymmetry and the Purely Deontic Asymmetry: under the former, both the Axiological and the Deontic Asymmetry are true, whereas under the latter, the Axiological Asymmetry is false. This means that it is easiest to split the evaluation of these views into two parts: first, we evaluate the Axiological Asymmetry, and then move on to the Deontic Asymmetry. Evaluating the former allows us to decide between the Full Asymmetry and the Purely Deontic Asymmetry. Evaluating the latter will then help us decide whether we should accept the remaining version of the Asymmetry or reject the Asymmetry altogether. So, let us begin by considering the axiological side of things.

It seems to me that defending the Axiological Asymmetry is not an easy task. First, note that if the Axiological Asymmetry is true, then no number of new happy lives can make the world better, and it does not matter whether we create a population of lives marginally worth living or a population enjoying unimaginable bliss. This seems to me absurd. Now, one might think that in claiming this to be absurd, I am simply assuming what needs to be argued. However, note that here we are not asking whether I have a duty or a moral reason to create happy people. Rather, the question is about which option is *better.* To push the intuition, you can imagine that you are forced (say, by God himself) to choose between two buttons, which will create either the marginal population or the extremely happy population. You must press one of the buttons, and the costs to you are the same either way. It seems implausible to me to deny that pressing the button creating the extremely happy population is the better option.

When it comes to betterness of options, the Axiological Asymmetry also creates the following counterintuitive result, originally pointed out by Broome (2004:147). Consider the following three options: 1) no one exists, 2) Ella exists with a moderately happy life, and 3) Ella exists with a very happy life. The Axiological Asymmetry implies that 1 is equally as good as 2, and 1 is equally as good as 3, which implies that 2 is equally as good as 3. But this seems implausible: 3 is better for Ella than 2 and worse for no one, so it seems obvious that 3 must be better than 2 simpliciter. Again, we might think that we have no duty to choose 3 rather than 2, but it seems much more difficult to claim that 3 is not better than 2.

Another reason, in my opinion, for being sceptical of the Axiological Asymmetry is that it seems to imply the following: to make things go best, a parent should organise things so that their child will be born (or otherwise enter the realm of the people who count) with a marginally positive life, so that they can then improve this life significantly. This follows from the fact that according to the Axiological Asymmetry, creating new happy lives does not add value to the world, but improving the lives of already-existing people (or people who will necessarily exist, and so on) does. If this were how axiological value is generated, then we would be able to 'play the system' by creating less happy people on purpose, so that we can then realise maximal value by helping them afterwards. For example, perhaps we should infect all babies with some disease that we can cure a few months after their birth. This implication is surely absurd.

For these reasons, I conclude that we should reject the Axiological Asymmetry, and with it the view I named the Full Asymmetry. In other words, if there is a version of the Asymmetry that we should accept, then it must be the Purely Deontic Asymmetry.

3.3.2. Is Deontic Asymmetry credible?

Consider next the Deontic Asymmetry. One natural argument for the view, in a broadly Scanlonian Contractualist spirit, goes as follows (Scanlon 1998, see also Horton 2021:486-489). If we do not create a happy life, there will be no one who can have a moral complaint against this choice. On the other hand, if we create a miserable life, or if we fail to improve the life of someone who already exists even though this would have been very easy for us to do, then there is someone who can have a complaint against our act. We have a moral reason to avoid outcomes that give someone a legitimate moral complaint against us. So, we have moral reason to avoid creating miserable lives and to improve already-existing lives, but we have no such reason to create happy lives.

I find this argument at least prima facie plausible. We can find something like the idea of legitimate moral complaints tracking wrongness in many forms of person-affecting ethics, as well as common-sense morality. For example, we do not typically think that the reason why we should not punch other people in the face is that this makes the world an overall less happy place. Rather, we think that punching someone would wrong that person, in a way that they can legitimately object to.

One important objection to this argument is that those who accept it may be forced to accept the conclusion of the *non-identity problem* (see Parfit 1984, Boonin 2014). To see this, consider the following case. Under Option 1, we create child A at welfare level 100, and under Option 2, we create child B with welfare level 2. Welfare level 1 represents the lowest amount of welfare where life is still worth living. Now, most people agree that if faced with this choice, one ought to choose Option 1. But notice that since each child only exists under one of the options and both have lives worth living, neither of them has a complaint against whichever choice you make. Even if you choose Option 2, child B has a life worth living, whereas child A never exists, meaning that neither of them has a complaint against your choice. So, the view I sketched above implies, counterintuitively, that both options are permissible.

Unfortunately, I do not have the space here to dive deeper into the debate revolving around the non-identity problem. For now, I want to make two points. First, I should say that there are some views that aim to avoid the non-identity problem, while holding onto the Asymmetry and the related person-affecting ideas (e.g. Frick 2020). Second, there are also authors who argue that despite the initial counterintuitiveness of the view, we should in fact accept the conclusion sketched above (e.g. Boonin 2014). For these reasons, the non-identity problem does not seem to me to be an immediate knock-down argument against Deontic Asymmetry. Furthermore, keeping in mind just how intuitive the Asymmetry itself is, I think that on balance, we at least cannot dismiss the Deontic Asymmetry in the same way that I dismissed the Axiological Asymmetry.

From this assessment, it follows that we must take seriously the possibility that the Purely Deontic Asymmetry is true. This, in turn, gives us a reason to reject strong longtermism, the strength of that reason depending on just how compelling we find that possibility.

Conclusion

In this chapter, I have clarified the ways in which the Asymmetry interacts with strong longtermism, paying close attention to how different interpretations of the Asymmetry along axiological and deontic dimensions can change the result. I have argued for three main conclusions.

First, I argued that accepting what I called the Full Asymmetry is not incompatible with accepting strong longtermism. Second, I argued that it is not possible to accept strong longtermism while also accepting what I called the Purely Deontic Asymmetry (PDA). Finally, I also briefly argued that we should take this form of the Asymmetry seriously. Much like anti-aggregative views, PDA shows us that not all plausible moral views will converge behind strong longtermism.

4

The Challenge from Fanaticism

In this chapter, I examine the argument that we should reject strong longtermism, because the assumptions that strong longtermism relies on imply *fanaticism.* Fanaticism, in this context, is the view that for any finite payoff v, there is some much larger finite payoff V, such that we should prefer the prospect of getting V with some very tiny probability $p > 0$ over getting v with certainty.[9] For reasons I will soon explain, many philosophers find this view objectionable. Via modus tollens style reasoning, this gives us a reason to reject the assumptions that ground strong longtermism. Consequently, unless we find some other way to justify strong longtermism that does not depend on these assumptions, we also have a reason to reject strong longtermism itself.

The essay has three sections. In section 1, I present the challenge that fanaticism poses to strong longtermism and explain why we should take it seriously. In section 2, I investigate whether existing literatures on decision theory and population axiology offer any promising ways to alter the assumptions that ground strong longtermism to avoid fanaticism. Here, I reject combining strong longtermism with either bounded axiological value or risk aversion but argue that there is a form of probability discounting that seems to provide a prima facie plausible solution, namely tail discounting. In section 3, I then defend the combination of strong longtermism and tail discounting. I conclude that, at least for those who see avoiding fanaticism as a necessary condition that any satisfactory view must meet, tail discounting does offer a way to do this while sustaining an interesting form of strong longtermism.

[9] For reasons of space and focus, I am setting aside the possibility of infinite value.

4.1. Fanaticism

4.1.1. What is fanaticism?

While the definition of fanaticism I gave above is somewhat informal, it should be sufficient for the purposes of this essay. For a refresher, here it is again.

> **Fanaticism:** for any finite payoff v, there is some much larger finite payoff V, such that we should prefer the prospect of getting V with some very tiny probability p > 0 over getting v with certainty.[10]

Assuming standard expected value theory (EVT), if V is large enough, then the expected value of this prospect will outweigh the (expected) value of getting v, no matter how close to zero probability p is. This judgement seems 'fanatical', for it implies that tiny probabilities of realising enormous value should dominate our decision-making. To keep things from getting convoluted, I will be using language that assumes positive value. However, I should note that, rather than realising positive value, we could also formulate fanaticism in terms of avoiding negative value. This would amount to avoiding even the tiniest probabilities of catastrophic outcomes at all costs.

Fanaticism, as I present it above, follows from the conjunction of two philosophical commitments. Firstly, as I already said above, we assume expected value theory (EVT) to be the correct theory of decision-making under uncertainty. According to EVT, one prospect is better than another if and only if it has the greater sum of probability-weighed value—in other words, the sum of the values of each possible outcome times their probability.

Secondly, arriving at fanaticism also requires that we accept an axiology that allows outcomes to be arbitrarily valuable with no upper (or lower) bound. Perhaps the simplest way to do this is to combine expected value theory with

[10] The talk about 'certainty' here follows the common way fanaticism is defined in the literature. We could, however, also frame the problem in terms of some very high credence short of 1. After all, we might think that barring logical truths, one should not have credence 1 on almost any proposition.

total utilitarianism, according to which the ranking of outcomes is determined simply by the total aggregate welfare of each outcome. But we could also, for example, arrive at fanaticism if we combined EVT with average utilitarianism, according to which the value of an outcome is just the average welfare of the population under that outcome. In this case, the very large payoff V could consist of, for example, one person living a blissful life for an astronomically long time.[11]

Having introduced fanaticism, I will now explain how accepting this view would lead to some very counterintuitive conclusions.

4.1.2. What is wrong with fanaticism?

To see why fanaticism is counterintuitive, consider the following case by Thomas and Beckstead (2021). You are on your deathbed, when a miracle happens: you are being offered one additional year of life by God. But then, the devil approaches you with an offer: would you change that one additional year to a 0.999 chance of 10 additional years? As a firm believer in EVT, you gladly accept. After all, the expected value of this deal is 9.99 years of additional life, almost ten times more than what God offered, and the decrease in probability is miniscule. The devil, however, is not done. He next offers you the chance to change your deal to a 0.998 chance of 100 more years, 0.997 chance of 1,000 years, and so on until you end up with a very tiny chance of a very large number of additional life years. As you accept fanaticism, you accept every offer the

[11] For the type of readers who find talk about *value* suspicious, it might be helpful to consider *preferences* instead. Under this framing, we might say that there is no upper bound on strength of preferences, meaning that some outcome might score arbitrarily high on desirability for that agent. The problem of fanaticism then becomes that if that kind of preference exists, then the agent should choose an arbitrarily low probability of satisfying such a preference over a certainty of satisfying any other preference whatsoever.

devil makes, leaving you in a situation where instead of a certain gain, you are now practically certain to die immediately.[12]

Fanaticism looks even worse, however, in moral cases, where others are affected. We might think that in the intra-personal case, the value of virtually any good we might acquire is naturally bounded above: after all, many of us have no wish to live forever. But it is much less plausible to think that there is some amount of total happiness or suffering after which the happiness or suffering of an additional person does not matter. This thought combined with EVT, however, commits us to counterintuitive moral choices. Consider, for example, the following case involving a choice between doing good in the present and aiming to influence the very long-run future:

> Laura has $100,000 to spend on improving the world. She has two options: 1) donating to the Against Malaria Foundation, or 2) funding a research project aimed at developing a mind uploading technology, which would allow morally valuable sentient life to exist indefinitely long into the future. Laura is certain that option 1) would result in 20 lives saved. However, she also has good reason to think that there is an extremely small but nonzero probability that her donation could lead to the success of the research project under option 2), which would translate to an unimaginably large number of additional, happy lives existing in the future. The expected value of option 2) beats the expected value of option 1), but due to the tiny probability of success under option 2), choosing it would almost certainly amount to simply leaving 20 people to die and gaining nothing.

[12] The above example, of course, assumes that we can place inherent, non-diminishing value on additional happy life. If this strikes you as a mistake, simply replace these goods with something you value inherently. At the very least, as I explain below, I believe you should do so when it comes to the happiness and suffering of other people. This is what truly matters for our purposes—the example above is merely for illustration.

If Laura accepts fanaticism, she will have to leave 20 people to die and fund the speculative, most likely futile research.[13] But this feels extremely difficult to accept.

4.1.3. Strong longtermism as a fanatical view

From the point of view of strong longtermism, the worry is the following. Greaves and MacAskill (2021:2) explicitly assume EVT and total utilitarianism as their starting point for strong longtermism. Unfortunately, these assumptions imply fanaticism, which in turn implies the deeply counterintuitive conclusions discussed above. These conclusions give us a reason to reject those assumptions, and consequently, reject strong longtermism.

However, while this is a serious worry, concluding that we should reject strong longtermism because of fanaticism would be too quick, even if one thinks that no plausible moral view can embrace fanaticism. This is because we may be able to modify the assumptions that strong longtermism relies on. Indeed, even Greaves and MacAskill (2021:17) think that strong longtermism is compatible with some degree of risk aversion, which represents a departure from standard EVT. Because of this, I take it to be an open question whether those in favour of strong longtermism are unavoidably committed to fanaticism.

In the remainder of this paper, I examine whether we can modify the assumptions that strong longtermism relies on in a way that avoids fanaticism. To be successful, these modifications should also be independently plausible and stay true to the original motivations for strong longtermism.

[13] While 20 lives is an actual lower bound estimate of AMF's impact, talking about certainty here is of course a simplification. But, as pointed out in footnote 10, we could just as well formulate the problem of fanaticism in terms of some sufficiently high credence short of 1.

4.2. Can strong longtermism avoid fanaticism?

4.2.1. Modifying strong longtermism: axiology

Fanaticism follows from EVT combined with any axiology that allows outcomes to be valuable with no upper (or lower) bound. Let us consider the possibility of modifying strong longtermism with regards to each of these elements in turn. I will begin with what I take to be the more straightforward possibility, namely bounded axiologies. Sections 2.2. through 2.5. will then focus on decision theory. For now, our question is whether strong longtermism is compatible with any axiology that places a bound on how intrinsically valuable outcomes can be.

To put it simply, it seems to me that the answer here is a resolute no. Strong longtermism explicitly builds on the idea that all our outcomes matter equally, no matter how far in time. But to have an upper bound on value would enable the situation that at some point, further consequences do not matter. The whole raison d'être of strong longtermism is to expand our moral concern to the very far future—this is done by rejecting any positive pure rate of time preference. Why would the strong longtermist first reject discounting the experiences of future people, and then put in place what seems like not mere discounting, but complete disregard of them? In short, placing an upper bound on value to avoid fanaticism seems to conflict with the original motivation for strong longtermism.

Even independently of strong longtermism, bounded axiologies seem very counterintuitive. First, note that such an axiology would make the best course of action now oddly dependent on facts about the far past: whether happiness and suffering matters now depends on, for example, how much total happiness the civilisation of ancient Egypt contained, because this fact in part defines how close to maximum possible value we are now. If ancient Egypt contained enough total happiness, there would be no value in us making better off the people who are alive today.[14] Second, note that we can generate fanatical conclusions by leveraging tiny probabilities of immense suffering, and so, to avoid fanaticism by modifying our axiology, we would need to place a lower bound on value. But to say that after some amount of suffering, the suffering of an additional person does not matter, is to cling to a cure that, in my opinion, is much worse than the disease.[15]

Finally, it is worth pointing out that even if we found some well justified upper and lower bounds on value, it would be very surprising if they just so happened to fall where the strong longtermist needs them to fall. This is because, on the one hand, if the bound is low enough to block fanaticism, it risks cutting out the possibility of astronomical value, meaning that the original argument for strong longtermism fails to go through. And on the other hand, if the bound is high enough for that argument to go through, then it risks not being able to avoid fanaticism. So, in addition to the arguments above, there is a great risk of ad-hocness involved in trying to save strong longtermism from fanaticism by this method. If there is a way to justify strong longtermism without committing to fanaticism, the answer must be found in decision theory instead.

4.2.2. Modifying strong longtermism: a plan for decision theory

Having discussed bounded axiologies and found them difficult to defend, I will now consider the second possibility I mentioned earlier, namely modifying the decision theory that strong longtermism relies on. I take there to be two prominent possibilities here: risk-aversion and discounting tiny probabilities. First, we might think that what is going on behind our rejection of fanaticism is some form of risk aversion: we are not willing to take the risk of zero payoff, in certain circumstances, even if there is the possibility of a very high payoff. For

[14] This *Egyptology Objection* is usually attributed to Parfit (1984:420), who raises it against average utilitarianism.

[15] Versions of these arguments also apply on axiological views where, strictly speaking, each additional unit of value matters, but the total amount of value approaches some limit.

example, with Laura's choice, donating to AMF guarantees that she saves at least some people, meaning she can be certain that her donation does not go to waste. Second, we might think that because letting tiny probabilities dominate one's decision-making feels counterintuitive, we should simply ignore such probabilities, or at least severely discount them. Again, Laura might think that the chances of success with the research project are so small that she should just discard the project and fund AMF instead.

In what follows, I will first consider the two most prominent decision theories that aim to incorporate risk aversion, namely expected utility theory (EUT) and Buchak's (2013) risk weighted expected utility theory (REUT). To cut straight to the conclusion, I believe that neither of these views can ultimately help those in favour of strong longtermism to avoid fanaticism. I will then consider the possibility of discounting small probabilities, concentrating on what Beckstead and Thomas call tail discounting. Again, to spoil the conclusion, tail discounting seems to me to be the most promising way of defending strong longtermism without committing to fanaticism. The view does have other costs, but these costs are potentially easier to accept than the costs of fanaticism.

Before all that, however, let us return to risk aversion. At least to me, explaining Laura's choice to donate to AMF with risk aversion seems intuitively promising. Furthermore, Greaves and MacAskill (2021:17) explicitly state that they believe some degree of risk aversion to be compatible with strong longtermism. I will begin by briefly considering what risk aversion is.

For our purposes, I believe the following intuitive idea from Buchak (2013) will suffice. According to Buchak, we can understand risk aversion as the preference for getting some good with value x for certain over taking a gamble whose mean value is x. For example, a risk-averse agent would prefer to take £10 for certain over a 50% chance of receiving £20 and a 50% chance of receiving nothing. In Buchak's terminology, this agent is specifically risk-averse in money. We can extend this definition: an agent is generally risk-averse in money if out of any two gambles with the same mean value, she prefers the one that is less spread out. For example, a risk-averse agent would prefer to take a coin flip where heads pays out £40 and tails £60 over a coin flip where heads pays out £10 and tails £90. In the context of strong longtermism, this could amount to de-emphasising the quest for astronomical amounts of value, and instead placing more weight on securing a palatable outcome with greater certainty. Note, however, that at this point we are simply talking about preferences of the agent making the decision, and not yet about functions that represent either utility or risk attitudes.

Having clarified this, we can now move on to decision theories that aim to incorporate risk aversion. I start with EUT.

4.2.3. From expected value to expected utility

Our first task with EUT is to be careful to distinguish between expected utility, as discussed in this subsection, and expected value, as discussed above. The value of an outcome, in our discussion, refers to an absolute, interpersonally commensurable measure of whatever we take to be inherently valuable, with some specified zero point. For example, a total utilitarian axiology tells us that the value of any state of affairs is the number of people times their average welfare, with the notion of a 'life worth living' providing the zero point on the scale. Unfortunately for our current purposes, philosophers sometimes simply equate the two, using the word 'utility' to refer to value (as in 'total utilitarianism'). In this chapter, however, we use the word 'utility' in a different sense, namely the one that arises from decision theory and economics.

Under EUT, 'utility' is a formal tool used to numerically represent an agent's preference-orderings and the relative strength of one preference in relation to another one. More specifically, when we talk about risk aversion, we are talking about cardinal utility, which is an interval-valued measure of an agent's preferences over some set of options. This utility function, however, is only unique up to a positive linear transformation, meaning that the choice of zero point, unlike with value, is arbitrary. This means that interpersonal comparisons of utility make no sense, since there is no such thing as absolute utility. Instead, the utilities in question are only defined in relation to other utilities, related to other preferences that the *same* agent has.

How exactly does EUT account for risk aversion, then? Under EUT, we assign each outcome a numerical utility, such that if outcome A has higher value than outcome B, then the utility of A, denoted by U(A), is also higher than the utility of outcome B, denoted by U(B). Beyond that, however, utilities need not straightforwardly follow values: indeed, this is what allows us to use utilities to model risk aversion. Under EUT, risk aversion corresponds to the assumption of diminishing marginal utility. To use a monetary gamble as an example, this means that each additional unit of the payoff of a gamble adds less utility than the previous one, and consequently, the utility function takes a concave shape. This can explain why an agent would prefer a certain gain of £10 over the coin flip for £0 or £20. If the first £10 adds much more utility than then next £10 after that, then the agent maximises expected utility by taking the certain option. For example, if U(£10) = 10u and U(£20) = 15u, then the expected utility of the gamble above is only 7.5u, compared to the 10u of the certain option. In other words, to be risk-averse with respect to X, according to EUT, is to have a concave utility function with respect to increasing amounts of X. In the above, X is money, but it could just as well be whatever we ultimately value in the moral sense, for example happy life years.

What does this have to do with fanaticism? Note that if we are very risk averse, then our utility function will be very concave, and if our utility function is very concave, then utility will have an upper bound. This should sound familiar: as with bounded axiologies, a bounded utility function gives us a way to avoid fanaticism. Maximising expected utility does not imply taking bets with tiny probabilities of immense value, for even that immense value is assigned the same utility as other lower-value, higher probability prospects.

Note that even though the structure of the argument is similar to the argument from bounded axiology, it is now not the axiology, but the decision theory that blocks fanaticism. We can combine EUT with, say, total utilitarianism, and allow outcomes to be arbitrarily valuable, while at the same time maintaining that when it comes to decision-making under uncertainty, a permissible way to evaluate prospects of even arbitrarily valuable outcomes involves a bounded utility function. In other words, we can grant that a world with an astronomical number of sentient lives spread across the stars would be better than a world where humanity goes extinct 10,000 years from now, while insisting that when it comes to decision-making under uncertainty, we can still prefer a certainty of the latter scenario over a small enough probability of achieving the former. This is simply to say that it is rationally (and morally) permissible to choose in a risk averse way, even at the cost of not maximising expected value.

Now, does this help those in favour of strong longtermism? I believe that the answer is no, for broadly three reasons. Firstly, there are some general, decision-theoretic reasons for thinking that EUT does not offer a satisfactory theory of risk aversion. Buchak (2013: chapter 1) surveys these, and I tend to agree with her analysis. Since the focus of this thesis is on moral philosophy, and I do not believe I can improve on Buchak's arguments, I will not go deeper into that topic here.

Secondly, coming back to our topic, it is difficult to find a utility function that avoids fanaticism by bounding utility above, while at the same time placing that bound high enough for the original argument for strong longtermism to go through. Under EUT, we can only avoid fanaticism if the utility function is bounded, meaning that the function must be very concave indeed. But this means that we are also very risk-averse, which in turn entails that we would often decline gambles with even favourable odds and take the certain option instead. Laura, for example, might judge that even a one in ten chance of saving a million people in the long run by funding medical research is not enough to overweigh the certainty of saving 20 people by donating to AMF. But such a high degree of risk aversion seems to run counter to much of the real-world, practical longtermist work. Now, one might of course disagree about the numbers, but the basic point is that finding a utility function that works for the strong longtermist, while at the same time avoiding fanaticism, is a vicious task.

Thirdly, I am also worried about the suitability of EUT for moral decision-making in particular. EUT has its natural home in economics, where ‘utility’ is usually understood as something like the subjective satisfaction that one gets from satisfying one’s preference. Under this view, risk aversion follows solely from preferences getting saturated: the fifth burger does not taste as good as the second one, so it makes no sense to risk your second burger in order to win three extra ones. But this picture fits badly with choices of moral importance. If every additional life saved is as good as the previous one in terms of moral value, then why exactly should my preferences not reflect that? It seems to me that EUT smuggles in the idea that it is permissible to not care about what ultimately matters, which looks like a moral mistake. Or, at the very least, this logic of explaining risk aversion does not seem to me to accurately capture the phenomenology of the kind of dilemma that we face with cases like Laura’s choice: we are attracted by donating to AMF not because we care about saving the first 20 lives more than the billions of lives after that, but rather because we fear that gambling with the research project may leave us unable to save anyone at all. [16]

So, in sum, replacing EVT with a risk-averse version of EUT does not seem to me to be a promising move, given both the problems involved in combining bounded utility with strong longtermism and those with EUT in general.

[16] A strict formalist interpretation of utility, according to which facts about a utility function amount to nothing more than facts about an agent’s preferences over gambles, might resist the idea of saturation. But such a strict formalist view seems to me to rob EUT of any action-guiding force, because under such a view, utility functions merely represent how the agent makes choices, rather than giving that agent guidance on how to make them. Compare this to an agent who first meticulously calculates the expected value of each option, and only then makes the choice based on that calculation.

However, many scholars think that EUT is not the best theory of decision-making under uncertainty available: we might instead prefer Buchak's (2013) risk-weighted expected utility theory (REUT). It is, therefore, worth checking if this theory could allow us to improve the situation for the strong longtermist.

4.2.4. From expected utility to risk-weighted expected utility

Under REUT, we rank uncertain prospects by transforming our utility function with a separate function r, which captures our attitudes to risk, and then maximise the risk-weighted expected utility. For example, for outcome y with probability p, the risk-weighted expected utility comes out as r(p)u(y). Function r has the following properties: $0 \leq r(p) \leq 1$ for all p; r(0) = 0; r(1) = 1; and r is increasing. For an illustrative example, consider the risk function $r(p) = p^2$ and assume the utility function for money to be simply u(£x) = x. Then the REU of receiving £10 with probability 1 is 10u, whereas the REU of a coin flip between receiving £5 and £15 is $0.5^2 * 5u + 0.5^2 * 15u = 1.25u + 3.75u = 5u$. This means that an agent with risk function $r(p) = p^2$ exhibits risk-averse behaviour: she will choose the sure payoff over the gamble with same mean value.

Having achieved a very basic grasp of REUT, we next need to work out whether it can be used to help strong longtermism avoid fanaticism. Unfortunately for the strong longtermist, this seems difficult. The first important thing to notice is that, as Thomas and Beckstead (2021) explain, if r is assumed to be strictly increasing, then risk-weighted expected utility can still grow arbitrarily large, unless the utility function itself is bounded above. This is because for any p > 0, r(p) > 0, meaning that if u(y) is arbitrarily large, then r(p)u(y) can also be arbitrarily large. This means that to avoid fanaticism, our decision theory must still adopt a bounded utility function, meaning that, for our purposes here, REUT makes no improvement to EUT. Moreover, this time the shape of the utility function cannot be explained by risk aversion, as attitudes to risk are already supposed to be handled by the risk function. In this case, as Beckstead and Thomas note, the most natural interpretation of the boundedness of utility function is that it is a bound on axiological value. I have already discarded this possibility.

Furthermore, there is a powerful recent argument from Pettigrew (2022), which shows that REUT, combined with a moderately risk-averse risk function and some other plausible assumptions related to longtermism, leads to a very counterintuitive conclusion. Pettigrew points out that making sure that there is a future and making sure that this future is a happy one are two very different tasks. This means that we may succeed in the former without succeeding in the latter. In other words, avoiding human extinction may bring about a very good future, but it may also bring about a very bad one. Now, if we are risk averse, or if morality requires us to be, then it seems like we do best by initiating extinction voluntarily. After all, this is the only certain way to avoid a future with astronomical suffering, and if we are risk averse, we care more about avoiding the very worst outcomes than achieving the very best ones. Voluntary extinction (with an appropriate calibration of credences) amounts to taking something like the safe outcome with certainty, and this is exactly what, according to REUT, risk-averse agents do. Note also that a similar argument can, at least in principle, be formulated for EUT as well, meaning that the problem generalises across risk-averse decision theories.

In summary, I believe that neither EUT nor REUT can help the strong longtermist to avoid fanaticism. Both these decision theories, together with a bounded utility function, seem to be either independently implausible, too weak to avoid fanaticism, or so strong that they contradict the original motivation and argument for strong longtermism.[17] For these reasons, I will now set risk aversion aside and move on to consider the possibility of discounting tiny probabilities.

[17] Note that it might still be the case that strong longtermism can accommodate some degree of risk aversion. My claim is merely that this seems unlikely to block the problem of fanaticism.

4.2.5. From risk-weighted expected utility to tail discounting

Though REUT is often presented in a form where r must be strictly increasing, Buchak herself also entertains, but does not ultimately endorse, the possibility that function r could be merely non-decreasing. Interestingly for our purposes, relaxing this assumption would provide a way to avoid fanaticism even with an unbounded utility function—which we could, given that it is unbounded, simply equate with a value function. This is because the assumption that function r is merely non-decreasing allows the possibility that $r(p) = 0$ for some range of p, for example $0 \leq p \leq 0.000001$. Essentially, this amounts to discarding all prospects with probability 0.0000001 and lower. Consequently, we can avoid fanaticism, because we may simply ignore very tiny probabilities of enormous value. This approach, even though it is here presented as a special case of REUT, is usually referred to simply as discounting tiny probabilities.[18]

Although capable of avoiding fanaticism, the type of probability discounting that I have entertained here is difficult to defend without some further modifications. Following Monton (2019), let us call the approach that simply ignores probabilities below some threshold *Nicolausian discounting*. As Beckstead and Thomas explain, this idea faces certain severe problems. For example, it makes the rational course of action very dependent on how outcomes are individuated, as we could adopt such a fine individuation that the probability of each outcome drops below the discounting threshold.[19] Furthermore, Nicolausian discounting violates dominance reasoning in a very counterintuitive way. The view ranks equally two lotteries that are identical, except that one of them pays out much more in certain possibilities that fall under the probability threshold—indeed, it could even be that these possibilities together add up to a significant probability, even though they all fall

[18] We could combine discounting tiny probabilities with EUT, stipulating that the expected utility of a prospect x is 0 whenever p is below the relevant threshold, and p(x)u(x) otherwise.
[19] This problem could, of course, be avoided if there was just one well-justified way to individuate outcomes. That seems to me unlikely, but here I am setting the possibility aside without argument.

below the threshold individually. But it seems obvious that the lottery with the added possibilities of extra payoff must be better.

For reasons of conciseness, I will not discuss these problems any further, but instead, skip straight to the solution that solves these problems, offered by Beckstead and Thomas. This solution is that instead of giving zero weight to outcomes below some probability threshold ε, we give infinitesimal weight to all outcomes that are extreme in the sense that they are in either the right or the left tail of outcomes. The tails are defined as follows: outcome x is in the left tail if the probability of getting x or worse is less than threshold ε and in the right tail if the probability of getting x or better is less than threshold ε. The view, then, is that prospect A is better than prospect B if and only if the expected value of A, conditional on getting a normal (middle) outcome, is higher than that of B, and in the case of ties, we use expected value conditional on getting an extreme (tail) outcome.

Tail discounting avoids the problems with Nicolausian discounting discussed above (see Beckstead and Thomas 2021:14, Kosonen 2022:245-248). Furthermore, tail discounting has many benefits over risk-aversion. Because tail discounting can block fanaticism without thoroughgoing, extreme risk aversion, it can still allow interventions aimed at improving the long-run future to have astronomical ex ante value, as long as the chances of success are above the relevant probability threshold, and there is a non-negligible chance of better and worse outcomes. While I will discuss this in more detail in the next section, this does not seem implausible: even if something like Laura's mind uploading research project ends up being discounted, there are longtermist interventions with better odds out there. Furthermore, given that tail discounting can be combined with risk neutrality (or some moderate risk-aversion) when it comes to middle outcomes, this approach also avoids Pettigrew's argument from risk aversion to voluntary extinction. Finally, the approach does not require a bounded axiology (or even bounded utility). Thus, tail discounting seems to be the most promising solution so far.

4.3. Is tail-discounting strong longtermism defensible?

4.3.1. Is tail discounting independently plausible?

I have now identified what I take to be the most promising way to justify strong longtermism without committing to fanaticism, namely tail discounting. However, while I have noted that this proposal avoids the most obvious problems with the other views I have considered so far, it is not yet clear whether it allows the original argument for strong longtermism to go through, and whether it is defensible in its own right. I will begin by considering the latter question first, and then move on to strong longtermism in section 3.2.

The biggest problem with tail discounting, I believe, is the way the theory works when we get close to the threshold ε. Beckstead and Thomas call this threshold timidity: for two payoffs x and y, it is never better to get x below the probability threshold than to get y above the threshold, no matter how much better x is than y, and no matter how close the two probabilities are. Furthermore, where to place this boundary is in a certain way obviously arbitrary: why should we go for, say, probability 0.000001 rather than probability 0.000002? It seems odd that an arbitrary choice like this should make such a drastic, categorical difference. And it also feels somewhat odd that we should make this choice in a way that happens to work for the strong longtermist.

However, while Beckstead and Thomas seem to take threshold timidity to be a sufficient reason for rejecting this view, I am inclined to think that the problem in question is not necessarily any worse than fanaticism, or any of the other ways in which fanaticism may be avoided. Beckstead and Thomas themselves show that to avoid fanaticism, we must either reject the transitivity of ‘better than’ when evaluating prospects or accept some form of timidity. Timidity, in its general form, means preferring some finite payoff x with probability p over any payoff y, no matter how large, with an arbitrarily close but slightly lower probability q. In addition to tail discounting, forms of timidity arise from all other ways of avoiding fanaticism that we have considered. For example, bounded utility functions produce this result if moving from payoff x to payoff y does not increase utility. Now, as I have already explained, I take tail discounting to be the most defensible of the options that entail timidity over fanaticism. Furthermore, I take accepting intransitivity to be a non-starter: to mention just one among many well-known arguments against intransitivity in decision theory, this option would leave us vulnerable to money pumps (see Steele and Stefansson 2020).

Faced with these difficulties, we might of course just embrace fanaticism, and contend that despite its counterintuitive appearance, it is correct. Indeed, it seems to me that this is what Greaves and MacAskill ultimately think. They suggest that there may in fact be no way for the strong longtermist to avoid fanaticism, but contend that given the difficulties with alternative positions, this is not a decisive argument against their view. In addition to Beckstead and Thomas' results, they also point us towards Wilkinson (2022), who outlines further costs that those rejecting fanaticism face.

I am inclined to think that while tail discounting is clearly better than intransitivity, it is unclear whether we should rather bite the bullet on threshold timidity or fanaticism. Luckily for the strong longtermist, however, the choice between fanaticism and threshold timidity ultimately does not matter in the current debate, as long as tail discounting itself is consistent with strong longtermism. To see this, note that if I am right about the untenability of the other options, then all of us must accept either fanaticism or threshold timidity. This means that if one objects to strong longtermism on the grounds that it unavoidably implies fanaticism, then she must accept threshold timidity, meaning that she should have no problem with tail discounting. And if she does not accept threshold timidity, then she must accept fanaticism, meaning that she has no reason to object to strong longtermism in the first place. Either way, strong longtermism is safe—unless, of course, tail discounting turns out to be incompatible with strong longtermism. This is what I turn to next.

4.3.2. Is tail discounting compatible with strong longtermism?

As I already noted, tail discounting does not seem to be in any obvious ways contradictory with the motivations for strong longtermism: it does not require the assumption that at some point further happiness or suffering does not matter, and it also does not commit us to anything that goes obviously against the spirit of strong longtermism, such as self-inflicted extinction. There is, however, one natural worry. Given that longtermist interventions often rely on leveraging tiny probabilities of immense value, it could be that tail discounting requires us to discount these interventions and focus on the short term, where the chances of success are better. If this was the case, then we would be forced to concede that there is no interesting form of strong longtermism that does not also commit us to fanaticism.

Some longtermists respond to this worry by maintaining that in many important cases, the probabilities that we are dealing with are above any reasonable discounting threshold. Ord (2020:167), for example, estimates that there is a 1 in 30 chance of an existential catastrophe occurring due to an engineered pandemic during the next 100 years. To me, it does not seem implausible that, with better regulation and more research in biosecurity, we could reduce this risk in a way that should not be discounted. For comparison, the risk of dying from a car crash during a whole lifetime in the US is less than 1 in 100. However, it is worth noting, as Greaves and MacAskill also admit, that this response seems much less plausible in the case of private philanthropy. This is because the chances of making a difference for any individual donation are much lower, and thus may well fall under the relevant probability threshold. The probability of an existential catastrophe is one thing, but the probability of some intervention, let alone a single donation, making a difference is quite another.

Those in favour of strong longtermism have, however, another response available. In the argument above, I have implicitly assumed that extinction from a pandemic is among the very worst outcomes possible. However, as Kosonen (2022:245-249) notes, tail discounting can leave even low probability outcomes intact, if these outcomes are not extreme in the sense that there is a non-negligible chance of both better and worse outcomes. Recall how tail discounting works: we give infinitesimal weight to any outcome x such that the probability of getting x *or worse* is below the threshold or the probability of getting x *or better* is below the threshold. Now, since many longtermist interventions are aimed at avoiding a premature human extinction, tail discounting will not discount the potential outcomes of these interventions if there is also a high enough possibility of outcomes better and worse than extinction. For example, in the left tail of the probability distribution we could have a perpetual global tyranny, and in the right tail we might have an astronomical number of digital minds, all experiencing immense happiness all the time. This would leave a variety of extinction events in the middle, unaffected by tail discounting.

Naturally, what I have said here is unavoidably speculative, and we might reasonably disagree about the probabilities that we should assign to the kind of scenarios discussed above. Adopting tail discounting to stop strong longtermism from implying fanaticism certainly raises the bar for longtermist interventions: these interventions must either have chances of success above the relevant threshold, or address outcomes that are not the very best or the very worst possible. However, thinking about the possibility of human extinction over the next few hundred years seems to me to imply that this bar is not unattainable. So, it seems to me that tail discounting does not block the argument for strong longtermism. Therefore, if we want to avoid fanaticism while holding on to strong longtermism, we should modify our decision theory to incorporate tail discounting.

Conclusion

In this chapter, I examined the argument that we ought to reject strong longtermism, because accepting strong longtermism would unavoidably commit us to accepting fanaticism. I focused on whether we could modify the assumptions that ground strong longtermism in a way that avoids fanaticism.

I began by discussing bounded axiologies and concluded that they seem hopeless. I then discussed two forms of decision theory that incorporate risk aversion, namely EUT and REUT, concluding that even though some versions of these theories may allow us to avoid fanaticism, the resulting views are either difficult to defend in their own right or unacceptable to those in favour of strong longtermism. Finally, I considered discounting tiny probabilities, and concluded that while Nicolausian discounting seems implausible, tail discounting provides the most promising way for the strong longtermist to avoid fanaticism. I pointed out that even if tail discounting faces severe problems, the view does not seem any less plausible than other ways of avoiding fanaticism, meaning that anyone objecting to strong longtermism because of its link with fanaticism should be happy to accept tail discounting.

In so far as tail discounting allows us to ignore the very most outlandish tiny probabilities of immense value, I believe it makes strong longtermism more intuitively plausible to many. But even while some of those in favour of strong longtermism may reasonably continue to hold on to fanaticism rather than accept tail discounting, the possibility of avoiding fanaticism with tail discounting is important to appreciate. This is because it implies that what Greaves and MacAskill (2021:25) believe to be “one of the most plausible ways in which the argument for strong longtermism might fail” need not keep those in favour of strong longtermism up at night.

5

Strong Longtermism as A Public Philosophy

Thus far, this thesis has presented three challenges to strong longtermism: anti-aggregative moral views, the procreative asymmetry, and fanaticism. While I take myself to have offered important insights into the interplay between these debates and strong longtermism, I have not yet offered much in terms of a positive view. This is what I turn to in this chapter. I will argue that strong longtermism is most defendable, and indeed an attractive view, when conceived as a *public philosophy*. The inspiration for this solution is due to Goodin's *Utilitarianism as a Public Philosophy* (1995).

The chapter consists of four sections. In Section 1, I present the basic idea of a public philosophy, and show, as an illustrative example, how Goodin uses it to defend a version of utilitarianism. In Sections 2 through 4, I then deploy this idea in defence of strong longtermism, addressing each of the three objections considered in the previous chapters of this thesis. Section 2 is on anti-aggregative moral views, Section 3 on the procreative asymmetry, and Section 4 on fanaticism and risk aversion.

5.1. What is a public philosophy?

5.1.1. The idea of a public philosophy

I begin by explicating the idea of a public philosophy in general. In the first few pages of his book, Goodin says the following:

> The thesis of this book is that at least one normative theory, utilitarianism, can be a good normative guide to public affairs without its necessarily being the best practical guide to personal conduct. (Goodin, 1995:4)

He later restates his focus in the following words:

> My concern in this book, true to the thrust of this introduction, is with utilitarianism as a public philosophy. My main concern is with the ways in which utilitarianism can be a good guide to public policies, without necessarily being a good guide to private conduct. (Goodin, 1995:26)

These quotes, I believe, contain the basic idea of a public philosophy. There is, however, one crucial ambiguity. Goodin often talks about utilitarianism being a good guide for public actors, which leaves open whether this is because he believes utilitarianism to be a *true* description of what public officials ought to do, or because utilitarianism is simply a good heuristic that approximates truth in an actionable way. In this chapter, I will opt for the former interpretation. This means that I reconstruct the idea as follows: some philosophical view can be understood as a public philosophy if and only if it produces true predictions concerning what public officials and collective, state-level agents ought to do, even if it fails to do the same for what individuals ought to do in their private lives.

One helpful way to approach the idea of a public philosophy, offered by Goodin, is to compare it to the way that different forms of utilitarianism have been individuated. The following three versions of utilitarianism might be familiar to the reader: act utilitarianism, rule utilitarianism, and motive utilitarianism. What distinguishes these views, at least according to Goodin, is what they posit as the level of evaluation: are we choosing between individual actions, more general rules, or the kind of motives we should have? In addition to this familiar distinction, however, we can also distinguish between who should do the maximising: each individual, or some larger collective? Here, understanding some view as a public philosophy amounts to placing emphasis on the latter option.

This general idea, I believe, can be helpfully demonstrated by Goodin's own account. Because of this, I will next offer a brief introduction to how Goodin defends utilitarianism as a public philosophy.

5.1.2. Utilitarianism as a public philosophy

According to (my interpretation of) Goodin, the situation of public officials is different from that of individual citizens in a way that makes utilitarianism a true description of what public officials ought to do, even if this is not the case for private citizens. The first, and in my opinion the most intuitively illuminating, example of this that Goodin offers is the importance of impersonality in utilitarianism. Utilitarianism has often been criticised for leaving no space for personal commitments to our own projects, principles, or special relationships, and thus eroding our integrity (Williams 1973). In the context of a public office, however, letting such personal commitments guide policy would be patently inappropriate. As Goodin (1995:9) puts it, "It is the essence of public service as such that public servants should serve the public at large. Public servants must not play favorites."

Overall, Goodin takes the situation of public officials to be different from that of individuals in three ways. First, public decision-makers have only limited knowledge available of each individual person's situation, meaning that they must instead rely on population-level data and an analysis of standard needs that almost everyone has, rather than tailor policy for each individual. Second, public officials must adopt rules and institutions that are fit for recurring situations over extended periods of time, unlike individuals who can change how they act more rapidly. Finally, related to the above, the decisions public officials make will by nature be public to everyone, meaning that these decisions must also be justifiable to everyone in a way that every private decision is not (see Goodin, 1995:63-65).[20]

These special conditions, in Goodin's view, give rise to a particular form of utilitarianism that has at its heart the notion that public officials should choose institutions, practices, and policies that best promote overall welfare. Goodin himself names this *government house utilitarianism*. Goodin conceives this as a form of rule utilitarianism applied by public, state-level actors. However, the ultimate justification of these rules is still meant to be welfare maximisation—it is the situation of public officials, as well as the need to solve collective actions problems, which force us to choose rules and policies rather than individual actions. This means that, contrary to how Goodin presents his view, it is best to think of the view as act utilitarianism applied to policies and institutions, instead of individual acts—after all, rule utilitarianism, strictly speaking, does not tell us to choose between rules, but instead tells us that an act is permissible if and only if it follows the set of rules whose universal acceptance would make things go best.

Goodin takes his government house utilitarianism to solve many of the potent criticisms of utilitarianism. We already saw how the context of public officials turns some potential vices of utilitarianism into virtues. To give a further example, Goodin argues that his view can avoid the objections that utilitarianism either demands too much of us by requiring self-sacrifice, or too little by allowing rights violations. With the former, the argument is, roughly, that placing certain demands on state level actors relieves individuals from overt demandingness. With the latter, on the other hand, Goodin thinks that the more rule-based structure of his view means that government house utilitarianism will avoid the most egregious violations of putative moral rights.

[20] Here, Goodin seems to have a very idealistic assumption, but I think it is close enough to truth to be useful. Even though the UK Government, for example, surely does many things that ordinarily citizens do not know about, the Government's actions are still under much more public scrutiny than the actions of most individuals.

The idea here, as I interpret it, is that, for example, a government that would violate the rule of law whenever this would seem to maximise the good would make life very unpredictable and stressful for its citizens. This would, according to Goodin's line of thought, produce a worse result than holding onto the rule of law.

This, of course, is a quick reading of Goodin, and I have not offered much of a defence of his claims. This, however, was never my ultimate intention—instead, the above is meant to illustrate the idea that I intend to use as a basis for a view I call *strong longtermism as a public philosophy*. This is what I turn to next.

5.1.3. Strong longtermism as a public philosophy

Having introduced the idea of a public philosophy and how Goodin uses this idea with respect to utilitarianism, I now want to propose a view I call strong longtermism as a public philosophy. I define this view as the conjunction of the following two claims.

> **Axiological strong longtermism as a public philosophy**: In a wide class of choices for social practices, public institutions and government policy, all options that are near-best overall are also near-best for the far future.
>
> **Deontic strong longtermism as a public philosophy**: In a wide class of choices for social practices, public institutions and government policy, we ought to choose an option that is near-best for the far future.

The basic argument for these claims follows Greaves and MacAskill's argument for strong longtermism simpliciter: given that future people matter and that there could be very many of them, the best we can do, and indeed what we ought to do in the public context, is to choose an option that makes the far future go best. Like Greaves and MacAskill's view, my view also talks about ex ante value and subjective oughts. The core difference is that, here, I take strong longtermism to be a more modest claim, in that it is only a true description of what public officials ought to when choosing between social practices, institutions and policies.

In what follows, I aim to show that the objections against strong longtermism that I have discussed in this thesis can be satisfactorily resolved when strong longtermism is understood as a public philosophy. Before that, however, one final point is in order. It might seem that the idea of a public philosophy that is significantly different from what we ought to do as private citizens entails a peculiarly fragmented account of morality. Therefore, it is worth pointing out some ways in which this division might be justified. I think there are at least three broad classes of views that might do the job.

First, under many moral views, it is trivially true that what one ought to do depends in an important way on the context. For example, consider a view which tells us that we should maximise happiness, except when this involves leaving one's children to perish. It is easy to see how something like this view would make something close to classical utilitarianism a true public philosophy, since setting public policy does not typically involve such a sacrifice of family members. Second, we might adopt a view that involves role obligations: for example, we might say that in the same way that parents have special duties to their children, public officials have duties to citizens. A public philosophy would then accurately describe the latter, even if it failed to capture other kinds of obligations. Finally, following the so-called political realists (see Rossi and Sleat 2014 for overview), we might go as far as saying that there is a distinct political normativity that is independent from moral normativity, and then say that a public philosophy captures the oughts of political normativity.[21]

It seems to me that Goodin's view falls somewhere between the first and second view mentioned above, given that he places emphasis on how the

[21] That said, my guess is that none of the contemporary political realists would in fact endorse strong longtermism as a public philosophy.

situation of public officials differs from that of private individuals: unlike the latter, public officials must keep in mind that public policies must apply to large groups of people in the same way, remain workable for extended periods of time, and be able to withstand public scrutiny. While I tend to think of role obligations as the most plausible route to a distinction between private and public morality, in what follows, I remain uncommitted on this question. Indeed, as will become clear in the next section, part of my argumentative strategy is to show that those advancing the criticisms of strong longtermism discussed in this thesis also need this distinction, meaning that at least part of the burden of proof for justifying it falls on my dialectical opponents.

5.2. Anti-aggregative views

5.2.1. Recapping the challenge

I will begin by considering the challenge from anti-aggregative moral views, discussed in chapter 2 of this thesis. Recall that under many popular, non-consequentialist views, it is *not* the case that when forced to choose between meeting competing moral claims, we should always meet the set of claims summing to the highest aggregate. For example, many philosophers think that if we can either save one person from dying or any number of people from breaking a finger, we should always save the person who stands to die, no matter how much total suffering the broken fingers add up to.

Now, this poses the following problem for the strong longtermist. Strong longtermism relies on the idea that the vast axiological stakes involved in how the long-run future goes means that we ought to set aside most competing reasons to do otherwise, and choose the option that makes the long-run future go best—this is, essentially, the Stakes argument. The argument, however, does not go through if we are not obliged to meet the moral claims of the group with the highest aggregate claim in the first place. If, instead, we are sometimes obliged to prioritise meeting the strongest individual claim instead, we might end up focusing on the very worst-off people among us in the present. This is because it is often possible to provide very major benefits to people currently in existence in a way that it is not for those living in the far future. In sum then, the challenge from anti-aggregative moral views is that if any such view is correct, then the argument for deontic strong longtermism does not go through.

There is an extensive recent literature debating the merits of anti-aggregative and fully aggregative views. While I am inclined to lean towards the fully aggregative side, I take this debate to still be inconclusive. Instead of trying to end that debate once and for all, my argumentative strategy here will resemble that employed by Goodin: I argue that while fully aggregative reasoning may or may not be the true view when it comes to individual morality, there are strong reasons to prefer fully aggregative views when it comes to state-level actors, because the situation of these actors is such that anti-aggregative views become implausible. Note the connection of this argument to the claim about the burden of proof for separating individual and public morality mentioned in the end of the previous section: what I claim here is that for those in favour of anti-aggregative views, there had better be some way of justifying this distinction, because the best that those in favour of these views can hope is that their view is a true description of what private individuals ought to do.

In the following two subsections, I offer two arguments for anti-aggregative views being implausible in the context of public actors. These relate to the situation these actors are in: they face a significant degree of uncertainty, their policies will be applied to a very large number of people repeatedly, and any policy they set will not be compared only to other policies in the same area, but all public spending. Individuals, on the other hand, often have a better idea of the direct impact of their choices, their choices are relatively limited in scope, and they cannot always transform one kind of resources into another.

5.2.2. Anti-aggregationism, risk and loss of life

Having set the scene, we can now move on to consider problems that arise from the uncertainty involved in setting any public policy and the ways in which anti-aggregative views struggle to deal with risk. As Horton (2020) shows, any partially aggregative view will face the following problem: such a view must either claim that 1) we should prevent one person from suffering a migraine rather than preventing a tiny chance of death being imposed on any number of people, no matter how large, or 2) what we ought to do in one isolated choice between a migraine and accepting some tiny risk of death is very different from what we ought to do in a sequence of such choices.

I am inclined to think that the best anti-aggregative views (which are all partially aggregative, rather than non-aggregative) should bite the bullet with 1). This follows naturally from what I take to be the core idea of any anti-aggregative view, namely that at least when the moral claims in questions are sufficiently different, one ought to satisfy the strongest individual claim, rather than any number of weaker claims. A certain migraine seems to me to ground an obviously stronger claim than being subjected to a tiny chance of death—as Horton notes, it is permissible to drive someone to a pharmacy, even at the cost of slightly increasing the risk of death for innocent pedestrians. This view, however, has the unfortunate implication that we ought to save one person from a migraine at the cost of, say, imposing on a billion people a one in a million risk of death, which, assuming the outcomes are independent, results in one thousand deaths in expectation. This is because for each individual person in the latter group, a one in a million chance of dying constitutes only a weak moral claim, one that is outweighed by the claim generated by a certain migraine. This problem can be generalised to any comparison between statistical and identifiable claims: for example, we can replace the migraine with a certain death and run a similar argument to the conclusion that we should sacrifice any number of statistical lives in order to save one identifiable life.

Having explained all this, I can now offer the first reason for why anti-aggregationist views seem implausible in the context of public officials. To see the problem, consider the following situation. You are on your way to a gig that would earn you £10,000, which you would use to save two statistical lives by donating to AMF. But now, you see a child drowning in a shallow pond. You decide to save the child, even though this means foregoing your gig and not being able to make the donation you were planning to do. This is not because of uncertainty, as you have full confidence in AMF's impact. Instead, you believe that this choice assigns an appropriate kind of priority to the identifiable life in front of you over the statistical lives you could have saved by donating.

It seems to me that most individuals might not face very many choices with this type of structure during their lifetime, and these choices do not affect a very large number of people. Consequently, the total number of extra lives lost due to prioritising identifiable victims is perhaps not so great that it makes the behaviour impermissible. However, things are different on the societal level. Take, for example, a policy where in order to respect the strongest moral claim, we save one person from a certain death rather than preventing a 0.1 chance of death befalling each of 15 people—meaning that we save one identifiable life over 1.5 statistical ones. Even such a moderate weighting, when applied to saving one million identifiable victims, leads to 500,000 excess deaths. At least to me, this is way too high a price to pay, even if we accept that satisfying the strongest moral claim carries some pro tanto moral importance. Now, given that public policy will affect an even greater number of people over many generations, setting policy in an anti-aggregative way would lead to an even greater number of lives lost. Therefore, I believe that even if anti-aggregative views might be permissible in some private interactions, public policy ought to be set in a fully aggregative manner.

One thing that increases my confidence in the claim that anti-aggregationists should accept 1) rather than 2) in Horton's dilemma is that some prominent proponents of such views seem to agree with me on this point. For example, Frick (2015) argues that Contractualists should prioritise meeting the ex ante weightiest moral claim, rather than the ex post weightiest claim. Importantly, Frick is aware that this view implies the problem above, namely that Contractualists should sometimes save an identifiable life over any number of statistical lives. To address this problem, Frick proposes a pluralist view where the strongest individual ex ante claim is only a part of what makes a choice morally right, with the effect of our choices on people's overall wellbeing supplying another important moral consideration. I think it is worth noting how such a pluralist view fits together with what I have said above: even if we accept the pluralist solution, it seems to me that the best this can do for anti-aggregative views is to salvage them in the context of private morality.

5.2.3. Anti-aggregationism and context

My second argument for rejecting anti-aggregationism as a public philosophy stems from a different consideration. The most plausible anti-aggregative views, I believe, are the ones known as deontic partially aggregative views. The most influential such view in recent literature is the one called Aggregate Relevant Claims (ARC), developed by Voorhoeve (2014). To simplify slightly, ARC tells us to satisfy the greatest sum of strength-weighted, relevant claims, where a claim is relevant if and only if it is sufficiently strong relative to the strongest claim with which it competes. The importance of the *relevance* of moral claims in this view, however, has the implication that what we should do depends in a sharp way on what the set of possible choices is like. As Tomlin (2017) has shown, this leads to some serious problems for anti-aggregative views, which I believe further supports my argument in this chapter.

To set the scene, consider first the following three-option case, adapted from Tomlin. We can save either A) one person from a severe harm, B) 10 people from a moderate harm, or C) 100 people from a minor harm, such that the moderate harms are relevant relative to the severe harm and the minor harms relevant relative to the moderate ones, but the minor harms are not relevant relative to the severe harm. According to ARC, we ought to choose option B). This is because the minor harms are not relevant to the strongest claim that they compete with, namely the severe harm, so they get discarded. After that, we compare options A) and B), and given that the sum of claims we satisfy by choosing B) is greater than that of A), we ought to choose B). Now, keeping this in mind, consider then a case where option A) becomes unavailable. ARC now tells us to choose option C) rather than B), because the sum of strength-weighted claims in the former is greater. But this is odd: how come removing an option that lost to option B) in pair-wise comparison can lead to making the people in group B) worse off? Intuitively, it seems that if B) is the best option in the set that includes {A, B, C} then it should also be the best option in the set only including {B, C}.

Some defenders of partially aggregative views are willing to bite this bullet, arguing that sometimes the set of choices we have makes an important moral difference. While I have some sympathy to this view, it seems to me that accepting such context-dependence has odd implications when it comes to public policy. As Tomlin points out, the partial aggregation literature usually only discusses cases where the competing claims cannot be jointly satisfied: we either save the group facing a severe harm, the group facing a moderate harm, or the group facing a mild harm, but never, for example, some of the people facing the severe harm and some of the people facing the mild harm. However, as Tomlin notes in the context of healthcare, this is not how public expenditure works. Ear-marking some expenditure to some specific treatments, let alone specific sectors of society, is always artificial in the sense that sovereign governments could, if they wanted to, move money from one area to another. We might, for example, increase the budget for education at the cost of decreasing the healthcare budget, or we could raise taxes to increase benefits.

What does this have to do with ARC? The point of all this is that when it comes to government-level policy, there is no way to individuate choice situations in the way that ARC requires, because in principle, any part of the public expenditure could be spent on anything whatsoever. To see this, consider the following case, again adapted from Tomlin.

> A public body has 1 million pounds. There are two people requiring life-saving treatments, each costing £500,000. There are also 20 young people who want to do a PhD, requiring a maintenance grant of £50,000 each. Which claims should be funded?

It is very natural to group the people requiring the lifesaving treatment into one group and the aspiring PhD students into another. But this is a mistake: there is no reason why we could not fund, for example, one life-saving treatment and 10 PhD students. As Tomlin puts it, no two individual claims in this case are such that it would be impossible to meet them both, so no two claims compete in the way that ARC requires. In fact, ARC, without further modifications, is silent on what we should do in a case like this. And as Tomlin shows, it is not easy to modify the view to accommodate case like this either. Thus, it seems to me that even if a private actor might face some cases that fit the structure of ARC, the view is not suitable for guiding public policy.

In addition to this problem, however, things look even worse for anti-aggregative views when we observe a related issue, also identified by Tomlin. Consider the following case:

> Stage 1: You can save either group A, which contains one person with a strong claim, or group B, which contains 10 people with moderate claims, such that the strength-weighted sum of claims is equal between these two groups.
>
> Stage 2: A single person with a minor claim is added to group A, and a million people with minor claims are added to group B.

The idea of this case is that because moderate claims are relevant in relation to the strong claim, it should be permissible to save either group at stage 1, and consequently, because many more claims are added to group B at stage 2, it should be permissible to choose group B. But ARC claims otherwise. Because the minor claims in group B compete against the strong claim in group A, they are irrelevant, whereas the minor claim in group A remains relevant as this claim only competes against the moderate claims in group B. So, according to ARC, we should save A.

Government policies almost always affect people in more than one way: for example, positive healthcare outcomes have all sorts of positive knock-on effects. Now, imagine that the strong claim in the previous case is to be saved from death, the moderate claim to be saved from paraplegia, and the minor claim is being able to achieve A-level qualifications (secondary school diploma). Imagine that the move from stage 1 to stage 2 represents us finding out that the people in both groups have children: the person in group A has one child and the people in group B have two each, totalling 20 children. We also find out that for each of these children, their parent being cured (and only that) allows them to continue with education. So, at stage 2, we add one minor claim to group A and 20 minor claims to group B. The result is that we are now obliged to save group A. But this seems to me unworkable: public policy cannot be so wildly sensitive to changes in people's circumstances. Imagine being a public officer telling the news to people in group B: "I'm sorry, but there is this person who has just had a child, so we're going to have to defund your care."

In this section, I have argued that even if anti-aggregative views have some plausibility in the context of private morality, they are not plausible when it comes to deciding what public actors ought to do. Consequently, even if it is the case that these views form an important objection to strong longtermism when applied on the individual level, they are not able to refute strong longtermism as a public philosophy. We can now move on to the procreative asymmetry.

5.3. Procreative asymmetry

5.3.1. Recapping the challenge

In Chapter 3, I argued that whether the procreative asymmetry is a problem for those in favour of strong longtermism depends on exactly what form of the asymmetry we adopt. I concluded that the form of the asymmetry that is both most problematic to strong longtermism and independently most plausible is what I called the Purely Deontic Asymmetry (PDA). This is the view that even though creating happy people makes the world axiologically better, this gives us no moral reason to create additional happy people.

The problem for strong longtermism is as follows. Simplifying slightly, strong longtermism tells us that we ought to do what makes the future go best. Given that, under PDA, creating happy people is a part of what makes things go best, strong longtermism implies that we ought to create happy people. But this is exactly the opposite of the deontic component of PDA, namely the claim that we have no moral reason to create happy people. So, it seems that accepting PDA would mean that we must reject strong longtermism.

5.3.2. Responding to the challenge

I believe that, as with anti-aggregative views, the problem that the procreation asymmetry poses for strong longtermism can be dealt with by framing strong longtermism as a public philosophy. Here, however, the argumentative strategy will be slightly different. With anti-aggregative views, I argued that these views are inappropriate for guiding public actors. When it comes to the procreation asymmetry, however, I want to suggest that the move to a public philosophy can allow us to construct a view which satisfies the most important intuitions motivating PDA. Here is the view: governments should guarantee that all citizens enjoy comprehensive reproductive rights and can freely choose to not procreate, while at the same time promoting procreation through generous parental leave, subsidised or free day care, and other such measures.

This view, I believe, provides a satisfactory response to the moral intuitions behind PDA. I take the two most salient motivations behind the deontic claim that we have no reason to procreate to be worries about demandingness and individual autonomy. A duty to have children would be incredibly demanding and, at least at the current state of technology, pose what seems like a threat to our bodily autonomy, as well as our freedom to make life plans that do not involve raising children. These implications, I believe, form a kind of reductio against any view that rejects the asymmetry, meaning that we are tempted to accept the asymmetry instead. However, my view responds to these worries by safeguarding everyone's freedom to decide whether to procreate, stripping away from PDA an important intuitive benefit of that view. Furthermore, the government promoting procreation in ways that do not infringe on anyone’s autonomy in such a profound way, nor place overtly demanding duties on anyone, seems like an appropriate way to recognise that creating happy people makes the world better. Recall that, importantly, PDA already accepts this axiological commitment, meaning that in making this claim, I am not introducing any new claim that my opponent does not already accept.

To make the intuitive idea behind my view clearer, I think we can find a useful comparison from a surprising direction: consider the case of militaries. In countries with no universal conscription, such as the UK or the US, people are drafted into the army on a voluntary basis, meaning that the state does not force anyone to join. Rather, enough people joining is guaranteed through incentives like money, career prospects and so on. This practice reflects the natural thought that while on the one hand, it is good for a state (or perhaps even obligatory) to maintain a sufficient army for public safety, the state should nevertheless refrain from coercing people into taking arms. In general, I propose, it can sometimes be required for a state to promote some end without coercing any individual to take action towards that end. Creating more happy people seems to me a plausible candidate for being one of those ends.

5.3.3. Responding to an objection

Now, there is a natural objection to my view which goes as follows. Even if we accept the general idea that governments should sometimes promote some end without coercing citizens towards it, it is still the case that PDA and strong longtermism are incompatible with each other. Furthermore, even if many people do in fact believe in the asymmetry due to reasons of demandingness and autonomy, there are also more sophisticated philosophical positions supporting PDA that do not rely on such claims. For example, one natural way to justify PDA is to adopt a person-affecting view where we have no reason to create happy people simply because failing to do so would not generate any moral complaints. After all, if a person does not come to exist, they cannot have a complaint against us. So, the kind of view I have suggested implies wasting resources on something that we simply have no moral reason to strive for.

This objection has an important point to it. My view will be most appealing for those who want to hold onto the asymmetry for the kind of reasons relating to autonomy and demandingness mentioned above, and less so for those who accept PDA due to person-affecting considerations. To this latter group, my response is that I believe the kind of strict person-affecting views sketched above to be either false, or at the very least, unable capture everything that morally matters. I have two reasons for this.

Firstly, it seems to me that purely axiological considerations can sometimes generate moral reasons. Consider the following, fanciful case:

> God kidnaps you and puts you into a room in heaven with two buttons. To be released, you must press one of them—there is no other way out. Button A changes nothing about the world, except for returning you to Earth. Button B also takes you back, but in addition, pressing it instantly creates a thriving, perfectly just and ecstatically happy civilisation on a distant planet. The choice makes absolutely no difference to your personal situation. Which button do you press?

If purely axiological considerations never generate moral reasons, it does not matter which button you press. After all, if you press button A, there is no one who comes to exist, so no one can have a complaint against that choice. If you press button B, the people who come to exist are as happy as anyone can be, so they would not complain either. This, to me, seems like an obvious mistake. If you can create a literal world of joy, at no cost to yourself, why would you not do it? Intuitively, you ought to press button B.

Secondly, it is worth mentioning that, at least in the absence of some further story, the kind of strict person-affecting view that might give rise to PDA involves biting the bullet with the non-identity problem. While this is unlikely to change the minds of those strongly attached to a view like this, it is still a cost worth mentioning.

On the other side of the equation, there are at least two important responses to these arguments. First, in response to me asking why, in the case above, we should not just go ahead and create the very good world, those in favour of a person-affecting view might instead ask: who would we be wronging if we did not do so? In my view, asking who would be wronged is simply not sufficient to capture all the moral reasons that we can have. Instead of just avoiding acts that generate complaints, I also believe that we have a moral reason to pursue actions that, lacking a more sophisticated expression, generate something like *gratitude*. But this reflects a fundamental disagreement that I do not, unfortunately, have the space to pursue further here.

Secondly, one might note that the state supporting procreation is different from my imagined case in the sense that, unlike the choice between two buttons, any public policy promoting procreation would cost money that could be used to do good elsewhere. I agree that this is an important point, but I believe all it shows is that a state should not use all its resources to fund such policies. The reasons that, in my view, are generated by the happiness of people we could create must be balanced off against reasons stemming from other considerations—indeed, in saying that states should protect procreative freedom, I have already made such a balancing act.

In summary, I have argued that public actors do have moral reasons to enact policies that lead to new happy people being created, at least as long as this does not involve coercing anyone to procreate. In this way, I believe the move to a public philosophy can make rejecting PDA more palatable for many, even though, as discussed above, it is unlikely to persuade everyone. Note that while I have discussed policies that involve supporting procreation, there are also other policies that lead to more people being created, namely those lowering the risk of human extinction. My claim is that even if something about acts that cause new happy people to exist means that individual citizens never have the obligation to perform them (say, because they would be too demanding, as with having children), states do have reasons to cause new happy people to be created.

5.4. Fanaticism and risk aversion

5.4.1. Challenge from risk aversion

Having discussed anti-aggregative views and the procreation asymmetry, our final task is to consider issues related to fanaticism. Here, unlike with the other two objections discussed in this thesis, I already offered a substantive solution: to avoid fanaticism, those in favour of strong longtermism could adopt a view called tail discounting. There is, however, still a further challenge that is linked with fanaticism.

Note that tail discounting does not tell us whether we should be risk averse in relation to the middle outcomes (the outcomes not in the tails). Recall also that, as we saw in chapter 4, a high enough rate of risk aversion could cause problems for strong longtermism. This is for two reasons: first, extreme risk aversion might mean that we ought to initiate a voluntary extinction to make sure that something even worse does not come about, and secondly, if risk aversion was understood as a bounded utility function of some kind, this could block the original argument from astronomical value of the future that strong longtermism relies on. In other words, strong longtermism still requires that we follow expected value theory (or something close enough) when it comes to the so-called middle outcomes for the long-term future, such as many extinction events.

Now, the potential problem for strong longtermism arises from the fact that many people do not actually behave like this. Instead, many of us are very risk averse in at least some areas of our life. This could mean that strong longtermism is still too ‘fanatical’ in the sense of allowing probabilities of immense value which are small, but not tiny enough to be discounted, to guide our decision-making to an extent that many people would object to. In other words, fanaticism might be much cheaper than often thought, in the sense of becoming a problem way before we get to the kind of probabilities commonly discussed in the literature.

5.4.2. States have much less reason for risk aversion

In response to the worry above, I argue that even if risk aversion is often appropriate for individuals, states should be broadly risk neutral. This follows the idea of strong longtermism as a public philosophy that runs through this chapter. The form of the argument here is similar to the case of anti-aggregative views: I argue that whether individuals ought to be risk neutral or not, risk neutrality is a virtue for state-level actors. I offer two arguments in favour of this view. Firstly, I believe the reasons that support risk aversion for individuals rarely apply to states. Secondly, adopting a risk averse policy on the state level leads to big losses in the long term.

Let us begin from the reasons that generally support risk aversion. Often, what makes us risk averse is our inability to absorb large losses: for example, a newly bought house burning down without insurance could leave one trapped in a life in debt. On the other hand, sometimes we might come across situations where we must choose between a certain gain or a small chance of a much bigger gain. Here, the catch is that if the chance of this large gain is sufficiently small, we might think that taking such bets in general is not a great strategy for making our lives go well, for our lives might be too short to take enough of these bets to make it likely that we get to enjoy the payoffs. This is essentially the reasoning that Monton (2019) puts forward, naming it the "YOLO argument" against expected utility (or value) theory. He explains:

> Maximizing expected utility can put one in a situation where one has a high probability of having one's life actually go badly. Because one only lives once, one has good reason to avoid choosing an action where one has a high probability of having one's life go badly, regardless of whether or not the action maximizes expected utility. (Monton, 2019:14)

It is easy to see how the combination of our inability to absorb losses and not having enough time to make sure maximising expected value pays off can justify risk aversion in many cases. For example, many of us would refrain from investing our pension savings in very risky ways, even if those risky prospects offered the maximal expected value. This is, firstly, because losing our pensions would be horrible, and secondly, because our few decades in regular employment do not offer enough time for these gambles to pay off.

In response to this, I maintain that neither of these reasons apply to state-level actors. Firstly, states are in an important way much more able to absorb losses than individuals, simply because states can divide up any losses across the whole population and over multiple generations. For many individuals (though not everyone), investing millions of pounds and losing it all would be devastating, whereas pretty much any state in the world can, and indeed does, make such investments all the time. Secondly, states 'live' much longer than most individuals, meaning that they also have much better chances of reaping in the benefits of maximising expected value.

5.4.3. Risk aversion leads to huge losses

The idea that states are better placed to reap the benefits from maximising expected value leads us nicely to my second argument for risk neutrality. This argument is simply that, in the long run, maximising expected value is a good strategy, because even for lotteries with low chances of winning, repeating them enough times means that the overall payoff averages out to the expected value with very high probability. On the flipside, this means that a state that consistently passes up bets with favourable expected value ends up losing out. For example, we can imagine a society that consistently takes chances to reduce the chance of extinction over very long timespans. Even if most of the interventions that get funded in this society did not make any difference, this policy would at some point prevent a major catastrophe, whereas the opposite policy would mean that sooner or later, such a catastrophe would take place. This seems to me to a very strong reason for states to be broadly risk neutral, at least when it comes to outcomes not in the tails.

In addition to thinking about time, we can combine this argument with the point that Greaves and MacAskill make, namely that sometimes what a society does is very different from what individuals do, because there are so many people involved. When enough people are involved, taking actions to avoid low-probability events with drastic consequences makes more sense, because more people means more iterations of the lottery in question.

Finally, the fact that risk aversion leads to huge losses when applied to large populations over long timescales also provides a response to one natural argument in favour of risk aversion. Pettigrew (2022) argues that when a group of people have differing risk attitudes concerning some choice we must make, the morally right course of action is to respect the preferences of the most risk-averse person in the group. Here, the recurring argumentative move to public philosophy becomes relevant: even if we accept that respecting the most risk averse individual generates a pro tanto moral reason, this reason is surely defeated by considerations of the large amounts of overall wellbeing that are at stake when we move from the individual to the societal level. In other words, even if respecting the most risk averse agent's preferences may or may not be the right thing to do when a small group of people hikes a mountain, it should not guide public policy.

Conclusion

In this chapter, I have provisionally argued that even if the three objections against strong longtermism discussed in this thesis pose serious problems to the view when applied to individual morality, they fail to refute strong longtermism as a public philosophy. I believe the argument is at its strongest when it comes to anti-aggregative views and issues related to fanaticism and risk aversion, and less so when it comes to the procreative asymmetry. Nevertheless, I believe that what I have said here is sufficient to at least make the case that we should take seriously the possibility of strong longtermism as a public philosophy.

Concluding Remarks

To conclude this thesis, let us look back to see what we have learned. After introducing strong longtermism in Chapter 1, I presented three objections to it, stemming from anti-aggregative views, procreation asymmetry, and fanaticism. I then proposed a potential solution, namely understanding strong longtermism as a public philosophy.

In Chapter 2, I argued that strong longtermism conflicts with partially aggregative and non-aggregative moral views, because what makes the far future go best might amount to meeting a very large number of moral claims that are, ex ante, not as weighty as the claims of the neediest alive today. This finding adds a new element of practical importance to the long-running aggregation debate in ethics.

In Chapter 3, I argued that what I called the Purely Deontic Asymmetry provides another serious objection to strong longtermism. This is because PDA implies that we have no moral reason to bring into existence happy people, even though it seems that this is necessary for making sure that the far future includes astronomical value—and is thus required, at least on some level, by strong longtermism.

In Chapter 4, I considered the objection that strong longtermism relies on implausible assumptions, because those assumptions imply fanaticism. I argued that those in favour of strong longtermism can avoid fanaticism by adopting tail discounting into their decision theory. This move comes with the significant cost of threshold timidity, but it seems to me that if one is determined to avoid fanaticism, then this is the best option that one has.

Finally, in Chapter 5, I suggested that even if we take the three objections discussed in this thesis seriously, we should accept a view I named strong longtermism as a public philosophy. This means that strong longtermism should be primarily understood as describing what states and public officials ought to do, rather than as a blueprint for individual morality.

As with any thesis, there are also many issues that I have not discussed. To name just one of them, I have completely set aside the idea that we might be clueless about the long-run consequences of our actions. Due to space constraints, I have also said very little about how strong longtermism interacts with the non-identity problem.

Yet, I believe that the breadth of issues considered in this thesis is sufficient to ground some general remarks. Recall that according to Greaves and MacAskill, their case for strong longtermism is fairly robust against variations in the underlying ethical assumptions. One way to understand the project pursued in this thesis is to see it as an extension of the moral sensitivity analysis that Greaves and MacAskill perform. So, what are the results?

If what I have said in the preceding chapters is correct, then we should push against Greaves and MacAskill's convergence claim on the level of individual morality. It seems to me that even leaving aside violations of serious side constraints and allowing many personal prerogatives, choosing an option that is near-best for the far future can be morally dubious for the three reasons discussed in Chapters 2-4. First, choosing such an option might involve impermissible types of interpersonal aggregation; second, it might require that we create happy people, even though we have no moral reason to do so; and third, such options may amount to betting on a tiny chance of immense value, such that we are almost certain to gain nothing.

There is a common thread running across these objections. Anti-aggregative views, the Purely Deontic Asymmetry, and the rejection of expected value theory in order to avoid fanaticism all involve accepting the idea that we sometimes have no moral reason to do what makes things go best (in expectation). Strong longtermism, on the other hand, relies on the idea that, at least when the axiological stakes are high enough, we do have a strong moral reason to do exactly that. It seems to me that this conflict originates from a fundamental difference between broadly non-consequentialist and consequentialist outlooks on ethics. Views placing emphasis on the former draw limits to the degree of convergence that strong longtermism can attract.

On the one hand, what this means is that the debate about strong longtermism is, in part, a debate about the extent that non-consequentialist views are plausible. On the other hand, however, it is also a debate about the proper realm of consequentialist reasoning. All plausible moral theories must accept that consequences matter—what needs to be decided is how much, and under what circumstances. This point, I believe, implies that there is a different kind of convergence to be found, which I articulated in Chapter 5. Even if we find that strong longtermism is not a suitable moral code for private interactions, I believe that we should nevertheless allow it a substantial role in guiding state-level policy.

All of this is, of course, theoretically interesting, and I hope that this thesis has made a modest contribution to the study of these issues. Beyond academic philosophy, however, there is an important practical lesson to be drawn. If what I say in this thesis is correct, then we, as a society, ought to do much more than we currently do to protect the long-term future of humanity. Whether it is the climate crisis, pandemics, advanced AI, or nuclear weapons, what we must do now is make sure that the history of humanity does not end before it has properly begun.

Bibliography

Bauman, Tobias, 2017. S-Risk FAQ. *Effective Altruism Forum*. Retrieved through https://forum.effectivealtruism.org/posts/MCfa6PaGoe6AaLPHR/s-risk-faq on 25/8/2022.

Beckstead, Nick, 2013. *On the overwhelming importance of shaping the far future*. PhD thesis. Rutgers University, New Brunswick.

Bostrom, Nick, 2003. Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, 15(3), 308-314.

Broome, John, 2004. *Weighing Lives*. Oxford: Clarendon Press.

Buchak, Lara, 2013. *Risk and Rationality*. Oxford: Oxford University Press.

David Boonin, 2014. *The Non-Identity Problem and the Ethics of Future People.* Oxford: Oxford University Press.

Frick, Johann, 2015. Contractualism and Social Risk. *Philosophy & Public Affairs*, 43(3), pp.175–223.

Frick, Johann, 2020. Conditional Reasons and the Procreation Asymmetry. *Philosophical Perspectives*, 34(1), pp. 53-87.

Goodin, Robert, 1995. *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Greaves, Hilary and MacAskill, William, 2021. The Case for Strong Longtermism. *GPI Working Paper* - No. 5-2021. Global Priorities Institute, University of Oxford.

Greaves, Hilary, 2017. Discounting For Public Policy: A Survey. *Economics and Philosophy*, 33(3), pp.391–439.

Horton, Joe, 2018. Always Aggregate. *Philosophy and Public Affairs* , 46(2), pp. 160-174.

Horton, Joe, 2020. Aggregation, Risk, and Reductio. *Ethics*, 130(4), pp. 514–529.

Horton, Joe, 2021. New and Improvable Lives. *The Journal of Philosophy*, 118(9), pp. 486-503.

Kamm, Frances Myrna, 1993. *Morality, Mortality: Death and Whom to Save From It.* Oxford: Oxford University Press.

Kosonen, Petra, 2022. *Tiny Probabilities of Vast Value*. DPhil thesis. University of Oxford.

MacAskill, William, 2022. *What We Owe the Future.* New York: Basic Books.

McMahan, J., 1981. Problems of Population Choice. *Ethics*, 92(1), pp. 96–127.

Mogensen, Andreas, 2019. Staking our future: deontic long-termism and the non-identity problem. *GPI Working Paper* - No. 9-2019. Global Priorities Institute, University of Oxford.

Monton, Bradley, 2019. How to Avoid Maximizing Expected Utility. *Philosophers' Imprint,* 19(18).

Narveson, Jan, 1973. Moral Problems of Population. *The Monist*, 57(1), pp. 62–86.

Nefsky, Julia, 2019. Collective harm and the inefficacy problem. *Philosophy Compass*, 14(4).

Ord, Toby, 2020. *The Precipice: Existential Risk and the Future of Humanity.* London: Bloomsbury.

Parfit, Derek, 1984. *Reasons and Persons*. Oxford: Oxford University Press

Parfit, Derek, 2011. *On What Matters, Volume II.* Oxford: Oxford University Press.

Pettigrew, Richard, 2022. Effective altruism, risk, and human extinction. *GPI Working Paper*, No. 2-2022. Global Priorities Institute, University of Oxford.

Roberts, Melinda, 2011a. An Asymmetry in the Ethics of Procreation. *Philosophy Compass*, 6(11), pp. 765–776.

Roberts, Melinda, 2011b. The Asymmetry: A Solution. *Theoria*, 77, pp. 333–367.

Roser, Max, 2022. Longtermism: The future is vast – what does this mean for our own life? Online article by *Our World in Data*. Retrieved through https://ourworldindata.org/longtermism on 29/08/2022.

Rossi, Enzo and Sleat, Matt, 2014. Realism in Normative Political Theory. *Philosophy Compass*, 9(10), pp. 689–701.

Scanlon, Thomas Michael, 1998. *What we owe to each other*. London: Belknap Press of Harvard University Press.

Schelling, Thomas C., 1968. The Life You Save May Not Be Your Own. In *Problems in Public Expenditure Analysis*, pp. 127-162. Edited by Samuel Chase. Washington: The Brookings Institution.

Singer, Peter, 2011. *Practical Ethics*. 3rd Edition. Cambridge: Cambridge University Press.

Steele, Katie and Stefánsson, Orri, 2020. Decision Theory. In *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), edited by Edward N. Zalta. Retrieved through https://plato.stanford.edu/archives/win2020/entries/decision-theory/

Tadros, Victor, 2019. Localized Restricted Aggregation. In *Oxford Studies in Political Philosophy Volume 5*. Oxford: Oxford University Press.

Taurek, John M., 1977. Should the Numbers Count? *Philosophy & Public Affairs*, 6(4), pp.293–316.

Thomas, Teruji and Beckstead, Nick, 2021. A Paradox of Tiny Probabilities and Enormous Values. *GPI Working Paper,* No. 7-2021. Global Priorities Institute, University of Oxford.

Thomson, Judith Jarvis, 1997. The Right and the Good. *The Journal of Philosophy*, 94(6), pp. 273-298.

Tomlin, Patrick, 2017. On Limited Aggregation. *Philosophy & Public Affairs*, 45(3), pp. 232–260.

Voorhoeve, Alex, 2014. How Should We Aggregate Competing Claims? *Ethics*, 125(1), pp. 64–87.

Wilkinson, Hayden, 2022. In Defense of Fanaticism. *Ethics,* 132(2), pp. 445-477.

Williams, Bernard, 1973. A Critique of Utilitarianism. In *Utilitarianism: For and Against*, pp. 77-150. Cambridge: Cambridge University Press.