Strong longtermism and the challenge from anti-aggregative moral views

Karri Heikkinen (University College London)

Global Priorities Institute | July, 2022

GPI Working Paper No. 5 - 2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Strong Longtermism and the Challenge from Anti-Aggregative Moral Views[1]

Karri Heikkinen

uctyklh@ucl.ac.uk

Abstract

Greaves and MacAskill (2019) argue for *strong longtermism*, according to which, in a wide class of decision situations, the option that is ex ante best, and the one we ex ante ought to choose, is the option that makes the very long-run future go best. One important aspect of their argument is the claim that strong longtermism is compatible with a wide range of ethical assumptions, including plausible non-consequentialist views. In this essay, I challenge this claim. I argue that strong longtermism is incompatible with a range of non-aggregative and partially aggregative moral views. Furthermore, I argue that the conflict between these views and strong longtermism is so deep that those in favour of strong longtermism are better off arguing against them, rather than trying to modify their own view. The upshot of this discussion is that strong longtermism is not as robust to plausible variations in underlying ethical assumptions as Greaves and MacAskill claim. In particular, the stand we take on interpersonal aggregation has important implications on whether making the future go as well as possible should be a global priority.

[1] This essay was originally written in spring 2021, before the publication of the updated version of Greaves and MacAskill's "The Case for Strong Longtermism" in June 2021. Because of this, I refer to the 2019 version of Greaves and MacAskill's paper throughout this essay. While the timing is unfortunate, I do not think anything important in my argument depends on which version of Greaves and MacAskill's paper I respond to. That said, the reader should note that some citations in this essay may not accurately represent the way Greaves and MacAskill now prefer to present their arguments. I am currently working towards an updated version of this essay, so any feedback would be greatly appreciated — please see the email address above. For comments I have already received, I thank Joe Horton, Dilara Küçük, and the audience at the UCL MPhil Thesis Preparation Seminar in October 2021.

Introduction

How much attention, resources and moral concern ought we devote to the very long-run future? Looking at the world around us, it seems that the generally accepted answer is: not very much. As Gonzalez-Ricoy and Gosseries (2016) illustrate, we live in a hegemony of *short-termism*. Our day-to-day politics rarely reaches even beyond the next electoral term. The 24-hour news cycle grabs our attention with a new topic every morning. In philanthropy, most donations go towards relieving the hardship of those alive now. And perhaps most strikingly, our carbon-intensive way of life risks leaving future generations with a radically impoverished planet.

Greaves and MacAskill (2019) believe that this is a grave moral error. Aiming to show that we should radically reorient towards the future, they argue for a view they call *strong longtermism*. The view consists of the following two claims.

> Axiological strong longtermism (AL): In a wide class of decision situations, the option that is *ex ante* best is contained in a fairly small subset of options whose *ex ante* effects on the very long-run future are best.
>
> Deontic strong longtermism (DL): In a wide class of decision situations, the option one ought, *ex ante*, to choose is contained in a fairly small subset of options whose *ex ante* effects on the very long-run future are best.
>
> (Greaves and MacAskill 2019:1)

From now on, I will refer to axiological strong longtermism as AL, deontic strong longtermism as DL, and the combination of these claims as strong longtermism.

In this essay, I argue that strong longtermism is incompatible with a range of non-aggregative and partially aggregative views in ethics. This result is important, because it shows that strong longtermism is not as robust to plausible variations in underlying ethical assumptions as Greaves and MacAskill argue. I believe that it also follows from this result that those in favour of strong

longtermism have to argue against non-aggregative and partially aggregative views, and adopt a fully aggregative view. This means that the side we take on the debate on interpersonal aggregation is likely to have important practical implications for whether we should spend our resources on improving the future or concentrate on the present.

Before going further, I should note that there are many other ways to challenge strong longtermism. Two particularly pressing potential problems for strong longtermism are the non-identity problem and the procreation asymmetry. However, I believe that these problems are relatively familiar to those working on strong longtermism. For example, Mogensen (2019) has explored the non-identity problem explicitly in relation to Greaves and MacAskill's paper, whereas Thomas (2019) has written about how the asymmetry might affect our moral attitudes towards the long-term future. Against this background, I believe that the challenge that non-aggregative and partially aggregative moral views pose to strong longtermism has been relatively neglected. The primary purpose of this paper, then, is to articulate this challenge and show why it deserves more attention than it has thus far received.

The essay has seven sections. In Section 1, I introduce the argument for strong longtermism. In Section 2, I briefly explain the variety of ethical views that different authors have taken on interpersonal aggregation. In Section 3, I present a hypothetical case to show that strong longtermism is at odds with non-aggregative and partially aggregative views, and I reject two initial attempts to dissolve this conflict. In Section 4, I present another type of case, highlighting how non-aggregative and partially aggregative views go against strong longtermism in circumstances involving risk. In Section 5, I locate the source of the issue at the so-called Stakes Principle that Greaves and MacAskill use in their argument. I also argue that those in favour of strong longtermism are better off simply arguing against non-aggregative and partially aggregative views, rather than trying to accommodate them. Finally, in Section 6, 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. Section 7 concludes.

1. Strong Longtermism

1.1. Clarifying the view

Let us begin by unpacking the two claims that make up strong longtermism, namely axiological strong longtermism (AL) and deontic strong longtermism (DL).

AL is a claim about axiology, meaning it is a claim about value. In this case, the claim is that the option that makes things go best, understood as realizing the most value possible, is simply the option that makes the long-run future go best. AL, however, 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 DL instead. According to DL, morality requires that we do what makes the long-run future go best. In slogan form, AL is about what is good and bad, and DL is about what is right and wrong. Notice that while Greaves and MacAskill defend both AL and DL, it could be that either one of them is true while the other is false.[2]

Both AL and DL are phrased in *ex ante* terms. Because determining the best action from the longtermist point of view involves a great deal of uncertainty, Greaves and MacAskill want to rule out evaluating the agent's actions *ex post*—that is, according to how things ultimately, as a matter of fact, turn out. What matters from the point of view of strong longtermism is choosing the right action from a set of uncertain prospects, based on the beliefs that the agent should have about the decision situation she faces. This same idea can be phrased in terms of subjective and objective oughts, the former corresponding with ex ante and the latter with ex post in Greaves and

[2] For example, we might think that even if AL is true, we are not obliged to do what is best, meaning that DL would be false. Or, one could think that even though AL is false, something like DL follows from, say, a duty to preserve the cultural achievements of the human kind.

MacAskill's vocabulary. 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 you this person actually has a severe aspirin allergy.

In practice, accepting strong longtermism would lead to a drastic change in how we spend our resources. According to Greaves and MacAskill, if strong longtermism is true, then for the purpose of determining what we should do, we can often simply ignore short-term effects and concentrate on the long run. This would imply spending much more than we currently do on things such as climate change mitigation and AI safety research. 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 recommend these interventions over more traditional ways of doing good in the present, such as donating to the global poor or volunteering in a local soup kitchen.

That said, strong longtermism is not intended to be an extremely demanding moral view, for it is limited to a "wide class of decision situations". This limitation is meant to allow that while in many cases the morally appropriate action depends on its long-run effects, other considerations can be decisive in certain other contexts, such as decisions involving one's family and friends. For Greaves and MacAskill, the paradigm case of a decision situation falling under the remit of strong longtermism is one where a cause-neutral philanthropist aims to use her resources to improve the world as much as she can. However, they think strong longtermism extends to many other situations beyond this paradigm, including governmental actors and individuals who aim to make socially impactful career choices.

1.2. Arguing for the view

Having defined strong longtermism, I now move on to explain Greaves and MacAskill's argument for the view. The authors begin by arguing for AL, using a combination of a plausible empirical assumption and an ethical claim. The empirical assumption is that the long-run future is vast, and thus involves an overwhelmingly large number of morally significant beings (Greaves and MacAskill 2019:4). The ethical claim is that all consequences of our actions matter equally, regardless of how far away in time these take place (Greaves and MacAskill 2019:5; see also Greaves 2017). Combining this claim with the assumption that the size of the long-run future is massive compared to the short term means that the amount of ex ante value we can produce by influencing the long term is larger than the amount of ex ante value we can bring about by influencing the short term.[3] Because of the vastness of the long-run future, this holds even when we take into account the greater uncertainty of the effects of our actions in the further future.

When it comes to DL, Greaves and MacAskill do not argue for this claim directly. Instead, after defending AL, they give an argument that tries to establish DL indirectly by using AL as a premise. Their argument goes as follows. For the first premise, note that after we have established AL, it seems that if AL is true, then it is likely to be true by a very large margin, again because the future is vast in expectation. From this it follows that whether we follow AL or not is a decision with massive axiological stakes. This is because every time we diverge from what AL recommends, we are potentially throwing away massive amounts of value, as the option that makes the long-run future go well is likely to be vastly better than the other options.

For the second premise, Greaves and MacAskill argue that the kind of non-consequentialist restrictions or prerogatives, which are usually taken to be the reasons for not choosing the option

[3] If this sounds complicated, the following analogy may be helpful. Imagine an archery competition where you get to shoot 1 million arrows, with the aim of scoring as many points as possible by hitting the target board. Given that each shot counts equally, it would be very unlikely that you score more points with your first 100 arrows than the last 999,900 arrows. Similarly, it would be unlikely that the short-run consequences of our actions generate more value than the long-run consequences.

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.

We can now see the argument for DL. First, whenever we face a choice between doing what makes the long-run future go best and some other option, we are facing a decision with massive axiological stakes. Secondly, non-consequentialist restrictions against doing what is best have no force when the stakes are big enough. So, our decision situation is such that non-consequentialist restrictions must be put aside. Therefore, what we overall ought to do in this decision situation is whatever makes the long-run future go best. Thus, we have arrived at DL. In essence, what happens here is that the Stakes Principle forces a convergence between AL and DL 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.

1.3. The overall project

Before moving on to aggregation, there are two further points relating to Greaves and MacAskill's overall project that the reader should be aware of. First, I should make clear that for Greaves and MacAskill, AL is more important than DL. This focus stems from the authors' commitment to effective altruism. In this context, we can understand effective altruism as a project which aims to identify how we can use our limited resources to do the most good possible (MacAskill 2019). Given this background, figuring out whether doing whatever makes things go best is strictly obligatory is not the top priority—rather, the most important task is to find out what in fact makes things go best. This means that while Greaves and MacAskill do provide an argument for DL, there is an important part of their overall project which does not necessarily collapse even if it

turns out that this argument fails. That said, as I will explain later, I believe the challenge I present in this paper is relevant even to committed effective altruists.

Second, it is useful to note that the argument laid out above forms only a small part of Greaves and MacAskill's paper. In the rest of the paper, the authors put considerable effort into defending the claim that strong longtermism is compatible with a wide range of philosophical assumptions, including different ethical views. This approach is motivated by the explicitly stated, practical goal of making strong longtermism appeal to a broad audience, so that we could eventually see a wider societal shift towards more longtermist thinking (Greaves and MacAskill 2019:3). For what it is worth, I admire this intention. Unfortunately, however, it seems to me that strong longtermism is not going to fit together with as wide a range of views as Greaves and MacAskill would hope. This is what I turn to in the next section.

2. Aggregation

2.1. Why aggregation?

My aim in this essay is to show that strong longtermism is incompatible with certain non-aggregative and partially aggregative moral views. My reasons for this focus are twofold. Firstly, it seems to me that those working on strong longtermism have not yet fully comprehended how certain aggregative views may pose a serious challenge to strong longtermism. In their paper, Greaves and MacAskill devote very little space to this issue. But even Mogensen (2019:11-12), who is critical of Greaves and MacAskill's argument due to reasons relating to the non-identity problem, maintains that aggregation does not form a serious problem for strong longtermism. As will soon become evident, I disagree with this evaluation.

Secondly, given that I believe there to be a conflict between strong longtermism and certain non-aggregative and partially aggregative views, I also believe that these views pose a serious

challenge to Greaves and MacAskill's ecumenical ambitions with regards to strong longtermism. If we want to push society in a more longtermist direction, then we need to understand how strong longtermism interacts with a range of popular non-consequentialist views, which take a critical stance on interpersonal aggregation. This is true regardless of whether one ultimately thinks that any such view is correct. And of course, in so far as non-aggregative and partially aggregative views can have some truth in them, figuring out their implications in the intergenerational context may have important practical implications.

2.2. Introducing aggregation

I find that the best way to get a grip of the different moral views on interpersonal aggregation is to consider the following situation. Imagine that 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? Different answers to this question allow us to helpfully categorise different views on aggregation.

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 claims

on both sides and see who comes out on top. It follows that even if harm x is much more serious than harm y, there always has to be *some* number of people suffering y that overweighs one person suffering x.

Finally, one can adopt a partially aggregative view (e.g. Kamm 1993; Scanlon 1998; Voorhoeve 2014; Tadros 2019). While a view of this kind can be specified in multiple ways, 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 larger number of people. 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.

Given the distinction between axiological and deontic versions of 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 often be interpreted in either deontic or axiological terms. We can think, in a purely axiological sense, that no number of minor harms or benefits can ever add up to enough (dis)value to overweigh the more serious harm or benefit. Alternatively, we can 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. For example, perhaps a certain moral ideal of respect towards other persons requires that we save the one from death rather than the many from a headache, even if we accept that in a strictly axiological sense this brings about a suboptimal outcome.

It is not possible to evaluate these views in much depth in the context of this essay. Generally speaking, non-aggregative views tend to seem plausible when harms x and y are very far apart in terms of seriousness, such as when comparing a death to many minor headaches. Conversely, fully

aggregative views seem attractive when the harms are only very slightly different. Partially aggregative views have the important benefit of fitting both these intuitions. However, these views also face severe problems in more complicated cases involving risk or repeated choices (Horton 2021). For the purposes of this essay, I want to 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.

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. For example, while some claims apply to anti-aggregative views in general, some only apply to deontic partially aggregative views. This terminology becomes particularly handy when we consider how those in favour of strong longtermism may reply to my arguments.

3. Lisa's Choice

3.1. The argument

We are now in a place to see the conflict between strong longtermism and anti-aggregative moral views. 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 enough to outweigh one death.[4] How about strong longtermism?

I think it is clear that if Lisa believes in strong longtermism, she will choose to set up the foundation. Firstly, given that strong longtermism depends on the idea that the future is vast in expectation, it is clear that the underlying idea here is to sum up the effects of our actions to the billions of people living in the future, in a fully aggregative manner. If we were not allowed to aggregate harms and benefits across all these people, 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 non-aggregative (and partially 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, positioning itself against anti-aggregative views.

[4] I assume throughout this paper that the aggregative views discussed are temporally neutral. One may reject this assumption, but to do that would be to raise a more fundamental objection against Greaves and MacAskill's central assumption that all the consequences of our actions matter equally regardless of how far in time they are. My aim in this essay is to articulate a problem that strong longtermism faces even if we accept this assumption.

3.2. Two initial objections

Greaves and MacAskill briefly note the possibility that anti-aggregative views may be in tension with strong longtermism, but they do not seem to be very worried about this. They point out that anti-aggregative views can only pose a problem for AL, which is the foundation of their overall argument and the more important one of the two claims they defend, if we interpret anti-aggregationism as an axiological view. They then note that the most plausible anti-aggregationist views take a deontic form, meaning that these views do not threaten AL in any way (Greaves and MacAskill 2019:17-18). Therefore, it may seem like the problem I raise here has little practical significance, especially when evaluated from an effective altruist perspective. After all, if AL is true, then regardless of whether DL 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 DL, still grants something like this line of thought.

I have two things to say about this. Firstly, note that AL alone would be much less controversial a claim than AL and DL together. The argument for the overwhelming moral importance of the long-run future follows rather naturally, if we assume utilitarianism from the outset (see Bostrom 2003). It is, therefore, worth investigating the more novel and controversial claim, namely DL. Secondly, I believe that the anti-aggregationist critique of strong longtermism is potentially much more damaging than those defending the view have realised. This is because deontic anti-aggregative views—the ones Greaves and MacAskill see as most plausible—require us to meet the strongest individual claim, regardless of what is best overall 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 a situation like this, AL would have no action-guiding significance, even if it was in a sense true. Thus, anti-aggregationism poses a serious challenge to both the truth of DL and the action-guiding power of AL.

Furthermore, it also calls into doubt the soundness of the argument from AL to DL. Therefore, I think those in favour of strong longtermism cannot afford to ignore anti-aggregative moral views.

As a different line of resistance, someone in favour of strong longtermism might object to my choice of example, as it diverges from the paradigm case of the cause-neutral philanthrophist that Greaves and MacAskill present. However, we should keep in mind that the authors themselves believe that strong longtermism will apply to many other situations beyond this paradigm. For example, they explicitly state that cause-neutrality is not essential from the point of view of strong longtermism (MacAskill and Greaves 2019:20). Therefore, I believe Lisa's choice should fall under the remit of the view.[5]

Ultimately, however, I also think that we should not get lost in the details of the example. The importance of Lisa's choice is to demonstrate a situation where we have to choose between delivering a very significant benefit to a small number of people in the present and setting up some mechanism which will reliably deliver smaller benefits to a very large number of future sentient beings in the long run. We can think of all sorts of candidates for what this reliable mechanism could be—for example, we might aim to promote broadly positive values or improve people's reasoning skills.[6] The important point is that in so far as strong longtermism instructs us to aggregate relatively small benefits over the very long term, this view is at odds with anti-aggregative ethical views. If the longtermist wishes to keep cases like this outside the scope of her view, then uncovering this restriction is an important result in itself.

[5] In the 2021 version of their paper, Greaves and MacAskill seem to limit the remit of strong longtermism more strictly, discussing mainly choices with no restriction to any cause area. This may mean that under the revised conception, Lisa's choice does not fall under the remit of strong longtermism. However, as I explain in the main text, the purpose of the case is to illustrate a point about aggregation, rather than act as a direct counterexample. Furthermore, as I suggest in Section 5.2., the continues existence of humanity itself may be analogous to Lisa's foundation: it is a mechanism that reliably brings a great number of moderate benefits that add up to potentially astronomical value in the long run.

[6] These, and other examples, were listed on 80,000 Hours' website in spring 2021, under "Our current list of especially pressing world problems". 80,000 Hours is an organisation with a significant emphasis on longtermism.

4. Sarah's Choice

4.1. 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, potentially even 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 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.

4.2. 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 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, because we will fail to reach herd immunity. Which program should we choose?

The problem with ex post anti-aggregative views is that they cannot distinguish between the two vaccination programs. This is because whichever vaccine we choose, 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 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.[7]

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 benefits in question are equally serious, then the

[7] 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.

certain benefit generates a more weighty moral claim than the uncertain one. And if we adopt an anti-aggregative view, then we ought to prioritise meeting the weightiest individual moral claim, even if there is a competing set of less serious claims that add up to a greater aggregate claim. 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.[8] Again, 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 the most plausible anti-aggregative views would tell Sarah to donate to GiveDirectly. So, again, we have a conflict between anti-aggregative views and strong longtermism.

Interestingly, the way anti-aggregative views can give some degree of extra weight to certain, identifiable benefits over statistical benefits[9] might allow us to explain the kind of 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 views implies that at least a part of the problem might fall under moral philosophy instead.

[8] 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. To avoid this issue, Frick (2015: 219-221) 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. This would allow us to maintain that the kind of pluralist anti-aggregative view that favours donating to GiveDirectly against what strong longtermism recommends is still a plausible view.
[9] By statistical benefits I mean the benefits we can predictably generate by applying some policy over a large number of people, but without us being able to know which of these people will end up receiving the benefits. The idea is exactly the same as the dichotomy of identified versus statistical lives coined by Schelling (1968).

5. Diagnosis

5.1. The source of the problem

We have now seen Greaves and MacAskill's argument for strong longtermism and how strong longtermism is in tension with anti-aggregative views. Specifically, I have identified two kinds of case in which these views are pitted against each other. First, there are choices between meeting a major moral claim in the present and setting up some mechanism which reliably delivers smaller benefits to a large number of people in the long run. Second, there can be choices between giving a small number of people a certain benefit in the present and giving a very large number of people a small chance of receiving somewhat similar benefits in the future. What should we make of this conflict?

To begin with, I think 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. Contrary to the Stakes Principle, these views imply that merely growing the aggregate stakes of the situation is not sufficient to necessitate the convergence of the best option and the option we ought to choose.

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 moral 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 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. Assume also that the number of broken nails averted would be so massive that this would add up to an overwhelmingly large aggregate harm. 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 all side constraints 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.[10]

5.2. 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. To tie this back to Lisa's choice, this restriction could mean that Lisa ought not to set up the foundation to treat minor ailments, but she should do so if the foundation can treat more serious illnesses instead—this would be to integrate a partially aggregative component into longtermism. Interestingly, the wider implication of this thought might be that we should pivot from unqualified longtermism towards caring more about so-called *suffering risks*—that is, aiming to make sure that the very worst kind of future scenarios never take place (Bauman 2017).[11]

[10] Some anti-aggregative views do deny axiology altogether, in which case the conflict with the Stakes Principle is even clearer. But as I show here, we need not go that far to get into problems.

[11] There is a further complication here relating to how we should understand giving benefits to future people, even if we were certain that those benefits will materialise. If we make the future go better, are we providing identifiable benefits to the specific people who in fact come to exist, or are we merely providing a vast amount of possible people a tiny chance of receiving the relevant benefit? Under the latter picture, we might think that even avoiding a

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). Personally, I think that this is ultimately the route that longtermists should prefer.

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 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. It is difficult to imagine how we could achieve anything like guaranteed major benefits in the far future—the bigger the benefit, the more uncertainty is typically involved. Given the amount of preventable suffering in the present and the uncertainty involving the long-run future, it seems unlikely that anti-aggregative views would converge with strong longtermism.

Of course, the picture I paint does not fully do justice to the fact that there are many difficult empirical questions in play here. Further research may turn out to suggest that the tension between anti-aggregative views and strong longtermism is not so severe after all. There is nothing in strong longtermism which would logically necessitate that all longtermist interventions must combine a large number of relatively small claims into high aggregate stakes. However, since the argument for strong longtermism relies on the idea that the vast *number* of future people tends to determine

future with intense suffering only provides statistical (albeit significant) benefits. This means that ex ante anti-aggregative views may end up disagreeing with strong longtermism even when it comes to suffering risks. I plan to explore this dynamic in an updated version of this paper.

the best course of action, rather than the idea that future people would somehow face more serious harms than those alive now, the conflict between strong longtermism and anti-aggregationism is not easy to dissolve.

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.[12] 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, the 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 very obvious fact that any intervention aimed at reducing existential risk involves betting on tiny 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 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.

[12] Even though I find this claim plausible, I anticipate that there are many people who claim that bringing someone into existence does not benefit them at all. However, denying this assumption does not help those in favour of strong longtermism. In short, if we cannot benefit people by bringing them into existence, then perhaps we only have comparatively weak reasons to make sure the human kind survives for a very long time. This leads us to issues related to the procreation asymmetry, which lie beyond the scope of this essay. The point of this passage is to show that anti-aggregative views pose a challenge for strong longtermism, even if we accept an assumption which is charitable towards strong longtermism.

The only tenable option for proponents of strong longtermism, then, is to argue against anti-aggregative views.

6. Are anti-aggregative views obviously implausible?

6.1. Greaves and MacAskill's objection

Before closing, 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 DL, the authors make the following brief remark concerning a possible objection against it:

> First, one might take a non-aggregationist view, and think that comparatively small benefits are not relevant to determining what one ought to do. […] Consider the example of someone alive in Britain during WWII, and considering whether or not to fight; or consider someone debating whether to vote in their country's general election; or someone who is deciding whether to join an important political protest; or someone who is reducing 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 person is question is obligated to undertake the relevant action. (Greaves and MacAskill 2019:24)

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 17 show that what they mean is both non-aggregative and *partially aggregative* views.[13] 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 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.

[13] This terminology is also present in the 2021 version of Greaves and MacAskill's paper.

6.2. A different problem?

In fact, it seems to me that, rather than responding to a potential objection, Greaves and MacAskill are raising a whole different moral problem here. This problem is 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 have to 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. Rather, the central idea is that we should aggregate the effects of an individual agent's actions over all those affected—think of the stylised cases where a doctor can either save one person from death or ten people from losing a limb.

Of course, we could try to introduce an account of how groups can act as moral agents, or how group membership is in some way morally significant. But as Nefsky (2019:3-4) shows, such accounts face some difficult problems. Furthermore, even if an account like this could be made to work, it is unclear whether the fact that the resulting view is fully aggregative would be doing any work in solving the inefficacy problem. Again, the interpersonal aggregation debate is typically understood to concern aggregation at the level of patients, rather than actors.

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.

7. Conclusion

In this essay, I have shown that strong longtermism is in conflict with non-aggregative and partially aggregative views. In particular, this conflict arises in at least two kinds of situations. Firstly, there are choices between giving a small number of people a major benefit in the present and setting up a mechanism that will reliably deliver a large number of relatively small benefits over the long term. Secondly, there are choices between providing a certain, major benefit to a relatively small number of people in the present and giving a tiny chance of a major benefit to a very large number of people in the future. Furthermore, it seems to me that we in fact face these kinds of choices when evaluating longtermist interventions, including projects aimed at reducing the risk of human extinction.

The most important implication of this result is that contrary to what Greaves and MacAskill argue, devoting a large share of our resources to make the long-run future go as well as possible is not something we should expect different moral views to converge on. 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 AL has a very limited relevance with regards to how we ought to spend our resources. This leaves a logical space for many non-consequentialists to reject strong longtermism.

If my argument is correct, then those in favour of strong longtermism cannot avoid the task of arguing against anti-aggregative ethical views. This finding adds a new element of practical importance to the long-running aggregation debate in ethics. It also calls for a more detailed investigation of just how big the rift between strong longtermism and anti-aggregative moral views is in practical contexts. For those of us who do care about the future, I hope that this short paper has demonstrated the importance of further work at this intersection.

Bibliography

Bauman, Tobias, 2017. S-Risk FAQ. *Effective Altruism Forum*. Retrieved through https://forum.effectivealtruism.org/posts/MCfa6PaGoe6AaLPHR/s-risk-faq on 25/4/2021.

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

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

González-Ricoy, Iñigo and Gosseries, Axel, 2016. Designing Institutions for Future Generations. In *Institutions For Future Generations*, pp. 3-21. Oxford: Oxford University Press.

Greaves, Hilary and MacAskill, William, 2019. The Case for Strong Longtermism. *GPI Working Paper* - No. 7-2019. 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. Partial aggregation in ethics. *Philosophy Compass*, 16(3).

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

MacAskill, William, 2019. The Definition of Effective Altruism. In *Effective Altruism: Philosophical* Issues, pp. 10-25. Edited by Hilary Greaves and Theron Pummer. Oxford: Oxford University Press.

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.

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

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.

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, 2019. The asymmetry, uncertainty, and the long term. *GPI Working Paper* - No. 11-2019. Global Priorities Institute, University of Oxford.

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.

80,000 Hours. Our current list of especially pressing world problems. Retrieved through https://80000hours.org/problem-profiles/ on 29/06/2021.