# Longtermism, Aggregation, and Catastrophic Risk

*Emma J. Curran*

## 1. Looking to the far-future

There is a growing number of philosophers who advocate focusing our attention on improving the prospects of those living in the very far future – say, one-thousand, ten-thousand, or even one-hundred-thousand years from now (Bostrom, 2003; Beckstead, 2013; Bostrom, 2013; Ord, 2020; Mogensen, 2020; Greaves and MacAskill, 2021). Advocates of *longtermism* point out that, in expectation, long-term interventions bring about far more good than short-term interventions, which tend to those living now or soon.

The view that we can bring about the most good by focusing on improving the prospects of those living in the distant future has been motivated, in large part, by the observation that humanity's future might be vast in size. Humankind's existence could, for example, be extended by colonizing other planets, which we could inhabit for up to 100 trillion years (Adams, 2008: 39; Beckstead, 2019: 82). Crucially, given the potential vastness of humanity's future, in expectation, the number of future persons to come will be enormous. Indeed, Hilary Greaves and Will MacAskill suggest that under a very conservative calculation, in expectation there are at least 100 trillion people to come, and that a more reasonable estimate would be at least one quadrillion (2021: 9). It is this observation which gives rise to the claim that those interventions which most improve the prospects of future people will bring about the most good. Consider:

> **Catastrophic Risk:** Arthur the philanthropist is interested in two possible interventions. With the money available to him, he could fund a medical treatment for ten patients. All of the patients have exhausted all other

> treatment options. If they do not have this procedure, they will all certainty die. The procedure has a high success rate, making it almost certain that all ten will survive. Arthur's other option is to put his money into artificial intelligence (AI) security research. Arthur has been advised that a country of one-hundred-million people are currently at a one-in-a-million risk of an AI-related fatal catastrophic event happening within their lifetimes. He has also been advised that his donation will reduce the risk of such an eventuality to five-in-ten-million.

In Catastrophic Risk, Arthur is faced with a choice. One option available to him is to treat a small number of people to substantially reduce their risk of a harm – that is, the group of ten. A second option is to reduce the risk of a harm for a vast number of other people by a very small amount – the group of one-hundred-million people. Despite the fact that by intervening on the large group, Arthur would only help each individual by a tiny amount, this would, in expectation, do a lot of good. Indeed, Arthur would save 50 people – 40 more lives than he would be investing in the treatment of the patients.

The choice between short-term and long-term interventions, in many cases, is very much like the choice faced in Catastrophic Risk between treatment and AI security research. Long-term interventions, like AI security research, generally only seem to improve the prospects of any far-future person by a tiny amount. This is due to the relative uncertainty with which we can shape the far-future; we're just not terribly good at bringing about far-future outcomes with much certainty. So, when it comes to long-term interventions, we can only ever raise the likelihood of bestowing a benefit to any future person by a tiny amount. On the other hand, at least some – if not a great many – of the short-term interventions available to us can substantially improve the prospects of their beneficiaries, given the relative certainty with which we can shape the present.

But long-term interventions clearly have something in their favour. Given the enormous number of individuals in the future who would receive a tiny increase in the chance of benefit, in expectation a *vast* amount of good would be bestowed by improving their prospects. This has two important implications. First, in expectation, the amount of good long-term interventions bring about is many orders of magnitude greater than the good brought about by short-term interventions. And, second, note that even the slightest increase to the prospects of future people generates a vast amount of good, in expectation. As such, long-term interventions which improve the prospects of future people by only a tiny amount more than alternative interventions are still vastly superior in terms of the amount of good they will bring about, in expectation.

So, it seems that longtermism provides us with reason to believe that we will bring about the most good, by an enormous margin, by investing in those interventions which most improve the prospects of future people.[1] Such a conclusion, longtermists argue, provide us with compelling, if not decisive, moral reason to choose such interventions. Indeed, given the enormity of the axiological stakes involved, some philosophers argue we have a moral obligation to invest in long-term interventions.[2]

If the conclusions of longtermists are correct, then our picture of beneficence ought to be dramatically revised. Longtermism might, for example, tell us to ignore those suffering around us in favour of attending to the needs of those who are yet to exist. This would be the case if those interventions which most improve the prospects of far-future people do not benefit individuals in the short term or do so to a significantly lesser extent than alternatives. And even if

1 Greaves and MacAskill call this claim, or something similar to it, "axiological strong longtermism" (2021: 6). Axiological strong longtermism has, however, been questioned, spawning a literature on expected utility theory and fanaticism, see: Bostrom, 2009; Balfour, 2020; Beckstead and Thomas, 2021; Wilkinson, 2022.

2 See, in particular, Greaves and MacAskill (2021: 27) who argue, through the stakes-sensitivity argument, that opportunity cost of not investing in long-term interventions generates an obligation to invest in them. Whilst not applied to the long-term context, complimentary arguments can be found from Horton (2017), Pummer (2016), and McMahan (2018).

the goals of long-term and short-term interventions align, such that long-term interventions also substantially help those in need now, longtermism would suggest that it was the value that such interventions brought about in the far-future which justified them, *not* their impact on those currently in need.

The argument of this paper might, to some, seem to resist these conclusions. Whilst this is possibly the case, my aim in writing this paper is to highlight a conflict between longtermism and, at least what I think is, plausible deontic scepticism about aggregation. I wish to demonstrate the scepticism about aggregation not only undermines an obligation to invest in long-term interventions, but also the permissibility of doing so. In §2 and 3, I outline concern about aggregation and the *ex-ante*/*ex-post* distinction, respectively. Once introduced, I proceed, in §4, to argue that from both an *ex-ante* and *ex-post* perspective, most long-term interventions will look unappealing to those who are sceptical about aggregation. And, finally, in §5, I discuss the implications of this incompatibility.

## 2. A problem of aggregation

Let us grant the claim that, in most cases, long-term interventions bring about far more good, in expectation, than available alternatives. Regardless, demonstrating that an intervention would bring about the most good, or the best outcome, is *not* the same as demonstrating that we have a moral obligation to choose that intervention. Moral philosophers are familiar with the many gaps between what would be best and what we are obligated to do. For illustration, consider:

> **Late Train:** A bystander is watching a large driver-less train when she notices an unconscious man lying on the track in the distance. The bystander happens to be standing next to a lever which acts as remote break for the train. She

can either pull the lever, causing all the passengers, of which there are N, to be very late, or allow the train to continue, fatally running over the man.[3]

Should the bystander stop the train? Intuitively, it seems impermissible for the bystander not to pull the lever. Indeed, no matter how fantastically big the train gets, and resultantly how large N becomes, it seems clear that we are committed to the impermissibility of allowing the train to drive over the man.[4]

Despite our intuitions in Late Train, there seems to be reason to favour running over the man. Namely, for some N, we would expect to bring about far more good by not stopping the train and allowing the man to be run over, as stopping the train would involve making a vast number of train passengers late. That is, we would prevent more harm by allowing the train to run over the man than we would by preventing the train from running him over. So, it seems in Late Train we have an instance in which the deontic comes apart from the axiological; even though it would best to allow the man to be ran over, we are morally obligated not to do so.

Where does such an assessment of Late Train go wrong? There are two diagnoses. The first is that in Late Train, we are wrong to claim that for some N, it would be better to run over the man. Such an account denies that small harms, like inconvenience, can ever aggregate together to be worse than a large harm, such as dying (Carlson, 2000; Dorsey, 2009; Temkin, 2012; Lazar and Lee-Stronach, 2019). Call this *axiological anti-aggregationism*. I will be placing axiological anti-aggregationism to the side for the remainder of this paper, instead choosing to focus on its deontic cousin.[5] Nonetheless, I believe much of what I say about

3 Another memorable example of this type is T.M. Scanlon's Transmitter Room (1998: 235).

4 One might wonder about the larger societal consequences of making a train late if said train is sufficiently large; for example, if enough people are late, the world economy might crash, or many people who needed urgent care might die. For the sake of the argument, let's presuppose that such catastrophes will not occur, and simply consider the cost to the individuals who are made slightly late.

5 I, in part, place axiological accounts to the side due to their apparent implausibility. See Horton (2021) for an overview of their problems.

the conflict between deontic anti-aggregationism and longtermism will apply, *mutatis mutandis*, to axiological anti-aggregationism.

The second diagnosis agrees that for some N, it would be better to run over the man. However, *deontic anti-aggregationism* points out that we arrive at the inappropriate deontic verdict because we aggregated the many complaints against the relatively insignificant harm of inconvenience such that they outweighed the complaint of the man against dying. If we did not permit such aggregation, then such verdicts would not arise. As such, in order to avoid morally inappropriate verdicts in cases like Late Train, some moral theorists are persuaded not to permit the aggregation of individual complaints.[6]

Scepticism about aggregation also comes from other sources, for example higher-level moral principles like the *separateness of persons* (Rawls, 1971: 26-27; Steuwer, 2020: 10-36). The thought is that given there is no person who will experience the sum of the many individual harms, then it does not make sense for all those harms to ground one aggregate complaint. To talk of one significant complaint against all the many harms of inconvenience in Late Train is to presuppose the existence of some entity who would experience all the many harms of inconvenience. But, of course, there is no entity like this – it is a fiction (Nozick, 1974: 32-22). A further consideration in favour of rejecting aggregation is that considering much smaller harms or complaints in the presence of morally significant ones is *disrespectful* to the person who stands to incur the significant harm; it seems to fail to take seriously the significance of what is at stake for them (Kamm, 1993: 144-63).[7]

6 Most famously contractualist follow this 'individualist restriction'. Importantly, contractualism does not also imply, or support, axiological scepticism, as contractualist moral theorising lacks an axiology due to the 'personal reasons restriction' (Kumar, 2003).

7 There is some debate, however, as to what respect requires. On one reading, respect requires us to not consider the weaker complaints at all, in the process of making the decision, whilst another reading simply states that it requires that the weaker complaints not form part of our reason not to satisfy the stronger complaints. See Mann (2022) for a discussion of this.

An individual has a complaint (or claim) in a decision situation if her well-being would be higher (or lower) under one of the available options.[8] The strength of an individual's complaint against a policy is a function of the impact the policy would have upon her wellbeing. Complaints compete with one another when they are not mutually satisfiable. *Aggregative* moral theories are those which make use of some choice rule which is applied to the aggregate sums of all the complaints which are satisfied under the various options. I will call moral theories which do not follow this *anti-aggregative*. Generally, there are two types of anti-aggregative moral theories: *fully non-aggregative* moral theories and *partially-aggregative* moral theories.

Fully non-aggregative moral theories do not permit any form of aggregation. Instead, they typically involve the decision-maker taking on the perspective of each person who has a claim in a decision-instance, and comparing its strength to each other competing complaint through a process of pairwise comparison (Scanlon, 1998). As a result, non-aggregative moral theories tend to require us to minimise the complaint of the individual who has the strongest complaint under the various options available (Kumar, 2003). It is crucial to note, however, that fully non-aggregative theories do not necessarily imply that one is not obligated to save the greatest number of people, if all the victims are facing equal harms (Taurek, 1977). It is possible to accommodate the intuition that we must save the greatest number without accepting some form of aggregation.[9]

Partially-aggregative moral theories tend to be similar to their non-aggregative cousins, except they permit aggregation in some instances (Voorhoeve, 2014, 2017; Lazar, 2018; Tadros, 2019; Rüger, 2020; Steuwer, 2021a; Mann, 2022).[10] Permitting aggregation of some complaints, whilst generally prohibiting it, has

8 Henceforth, I will simply talk of complaints.
9 For discussion, see Kamm (1993, 101-119), Scanlon (1998: 232-33), and Otsuka (2006).
10 Strictly speaking, Steuwer's account is non-aggregative, as rather than aggregating 'relevant' claims, he allows them to balance against each other (2021a). Nonetheless, what I say about partially-aggregative moral theories applies equally to his account, given the account of relevance he sketches. This also applies to Tadros (2019), alongside van Gils and Tomlin (2020).

the benefit of both respecting our convictions in cases like Late Train, whilst also cohering with other ground-level moral intuitions. For example, there are cases in which it seems to be appropriate for complaints against lesser harms to aggregate against greater ones; take, for example, the decision between saving ten people from quadriplegia and one person from death.

There are numerous proposals for how to best specify a partially-aggregative account. However, the argument I wish to make in this paper, I believe, applies to all such accounts of which I am aware. As such, I will only make use of a very broad sketch of partially-aggregative accounts, as opposed to any particular one. Generally, partially-aggregative accounts tell us to choose the option which satisfies the greatest sum of strength-weighted, relevant complaints.[11] A complaint against a harm x, is relevant to a competing complaint against a harm, y, if and only if the former is sufficiently strong relative to the latter.

## 3. How to evaluate claims under risk

Anti-aggregationism seems an obvious way to undermine the case for longtermism, as longtermism seems to hang upon the moral reasons that vast axiological stakes supply. But, before we begin to assess long-term interventions from an anti-aggregationist perspective, I must address another level of complexity.

Recall, long-term interventions are risky in nature; when we engage in them, we do not have any certainty about whether they will bestow a benefit on any future person. When evaluating risky choices, we can adopt two different perspectives: the *ex-ante* perspective and the *ex-post*. The distinction between *ex-ante* and *ex-post* has traditionally been characterised as a choice between focusing on an intervention's impact on individuals' *prospects* or its impact on group-level

11 The initial proponent of this general type of view is Alex Voorhoeve, who calls it "Aggregate Relevant Complaints" (2014; 2017). Not all accounts tell us to do precisely this, but to my knowledge the difference in their process isn't significant for the cases I discuss.

*outcomes*. In this sense, the choice between *ex-ante* and *ex-post* is sometimes seen as a distinction between the sorts of goods under distribution.

Another way of illuminating the distinction is as a choice regarding the nature of the individuals who claims we attend to. When we assess decisions *ex-ante*, we are considering the claims of individuals whose designation is independent of the outcome, such that regardless of what the outcome is – for example, who ends up bearing the costs or the benefits of an intervention – the designator will pick out the same individual. Examples of ways of designating individual's *ex-ante* include the use of legal names, some personal information, or, in the case of Catastrophic Risk, 'person one', 'person two' and so on until 'person 100,000,000'. We then calculate the claims each of these people have for a policy by calculating their expected interest in it. We do so by taking the difference between their expected well-being given the policy was implement and given the policy was not implemented.

On this account of the distinction, *ex-post* we are considering the claims of individuals whose designation is dependent on the outcomes of an intervention.[12] As such, the designator will pick out different people depending on what the actual outcome is. Whilst there are many ways to do so, *ex-post* analyses tend to use facts about the distribution of outcomes given a policy to list the individuals affected. One way to do so, as suggested by Bastian Steuwer, is to rank individuals, such as: 'the worst off given policy X', 'the second worst off given policy X', and so on until 'the least worst off given Policy X' (2021b: 118). Other plausible ways to designate people involve picking out normatively important outcomes, including 'the person who would die if not treated' or 'the loser of the policy', and so on.

In attending to the claims of people whose designation is dependent on the expected pattern of outcomes, *ex-post* moral theories focus on the chance that

12 As such, the distinction between *ex-ante* and *ex-post* often, though not always, tracks the distinction between designating people rigidly and non-rigidly, see Steuwer (2021b: 118).

*someone* will incur a harm, not that any particular person will (Otsuka, 2015: 86-89). As such, we discount each of these complaints by the difference the intervention would make to their improbability of occurring.[13]

To get a clearer sense of how we calculate claims *ex-ante* and *ex-post* and also how the two perspective can come apart, consider the following cases:

> **Dose Distribution**. We have five doses of a medicine for a deadly disease. Bernard has the disease. If we give him the medicine, he will almost certainly recover to full health, if we do not give him the medicine, he will almost certainly die. Caspar, Donald, Elizabeth, Frances, and Gerald are each at risk of developing the disease, with certainty that exactly one will develop it, but we do not know which. We can vaccinate each of the five against the disease using one dose of the medicine, making it such that no one will develop the disease. How do we distribute the medicine doses?[14]

What are the claims for the medicine *ex-ante*? Well, *ex-ante*, Bernard has a complaint of death in favour of the intervention which would give him all five doses of the medicine – without the medicine, he will almost certainly die, and with it he will almost certainly live. In favour of the vaccine intervention, each of the five has a complaint equal to how much the intervention would improve their prospects, namely by a one-in-five risk of death.

*Ex-post*, the complaints look a little different. The complaints against the vaccine intervention can be generated by listing all the individuals present in the

13 We do this, in part, to prevent ex-post assessments from being overly risk-sensitive. Take, for example, a decision to create a particular flight path. Let's say permitting this path creates the small risk that a plane will crash over a small island and kill one of the inhabitants. If we did not discount complaints ex-post, then the complaint of death from the islander would clearly be more significant than any claim of convenience we had. Rather, we must discount the complaint of death by the difference that taking this path makes to the improbability of a dearth occurring. Given the difference is sufficiently small, the ex-post complaint won't preclude air travel. See Otsuka (2015) for discussion.

14 This case is paraphrased from Norman Daniels (2015: 118).

case from worst-off the best-off. The worst-off if the vaccine is given – Bernard – has a complaint of death. The rest of the individuals do not have a complaint, as they are not harmed by the vaccine interventions. The claims in favour of the vaccine intervention can likewise we generated, with the worst-off member of the group – the person who would die if not given the vaccine – also having an *ex-post* complaint of death.

This case should show that the distinction between *ex-ante* and *ex-post* is not an ethically unimportant one; on an *ex-ante* assessment, the complaints seem to decisively favour the treatment option, whilst on an *ex-post* assessment, considerations of complaints seems to treat the treatment and vaccine options as identical. The *ex-post* perspective does justice to the intuitive idea that what matters, morally, is whether *someone* would die. That we cannot in advance point to a particular member of the group of five whose life we'd save is irrelevant, so long as we know we would, indeed, be saving a life.

To further illustrate the process of evaluating *ex-post* complaints, consider a variant of Dose Distribution in which we know that if left unvaccinated there is a sixty-percent chance of exactly one member of the group of five developing the illness. In this case, the *ex-post* claim in favour of vaccinating the group of five would be the complaint of death from the person who would die otherwise, discounted by the difference the intervention made to the improbability of it occurring, namely discounted by forty-percent

Finally, consider a variant of Dose Distribution in which the group of five each had an independent one-in-five risk of death. In this case, there are five *ex-post* claims in favour of the vaccination option, as five people could possibly die if left untreated. However, the probability of each of these deaths occurring follows a binomial distribution. As such, each of these five claims is discounted at a different rate, with the first complaint of the person who would die being

discounted by the likelihood that one person would die, the second complaint discounted by the likelihood that a second person would die, and so on.

With the distinction between *ex-ante* and *ex-post* moral theories at hand, we can now outline the four types of anti-aggregationist moral theory: *ex-ante* fully non-aggregationism, *ex-ante* partial-aggregationism, *ex-post* fully non-aggregationism and *ex-post* partial-aggregationism. In the following section, I demonstrate that all four of these are in conflict with longtermism in important ways. In particular, I demonstrate that *ex-ante* anti-aggregationist moral theories will judge the claims generated by long-term intervention to be particularly weak, making long-term interventions unappealing, especially in comparison to available short-term ones. I then show that whilst *ex-post* anti-aggregationist moral theories permit us to invest in some long-term interventions, they prohibit investing in a large and important class of longtermist activities.

## 4. Risk, aggregation, and catastrophes

With the two distinctions at hand, we are now in a position to evaluate long-term interventions from the perspective of anti-aggregationist moral theories. Recall our analogue for this decision:

> **Catastrophic Risk:** Arthur the philanthropist is interested in two possible interventions. With the money available to him, he could fund a medical treatment for ten patients. All of the patients have exhausted all other treatment options. If they do not have this procedure, they will all certainty die. The procedure has a high success rate, making it almost certain that all ten will survive. Arthur's other option is to put his money into AI security research. Arthur has been advised that a country of one-hundred-million people are currently at a one-in-a-million risk of an AI-related fatal catastrophic event happening within their lifetimes. He has also been advised

that his donation will reduce the risk of such an eventuality to five-in-ten-million.

Let's begin by assessing Catastrophic Risk *ex-ante*. How would an *ex-ante* fully non-aggregationist interpret the claims in this case? Well, *ex-ante,* each of the ten's complaints seems significantly greater than any of the one-hundred-million's. *Ex-ante*, each of the ten can claim that if not given the treatment they will certainly die, whilst if they are given the treatment they will, *almost* certainly, survive. On the other hand, any of the one-hundred-million can only claim that they will be exposed to a one-in-a-million risk of death if not treated. Comparing each of their individual complaints against each of the ten's complaints, it seems that *ex-ante*, fully non-aggregationist moral theories will favour treating the ten – that is, they will favour the short-term intervention analogue. In fact, they would claim that it was *obligatory* to invest in the treatment option, if you were to invest in either, as investing in the long-term intervention would be unjustifiable.

What of the *ex-ante* partial-aggregationist? They will come to the same conclusion unless they believe that complaints against small risks of a harm are relevant to complaints against large risks of that harm, such that, *ex-ante*, we can permissibly aggregate them against one another.[15] I find this, however, implausible.

Whilst there is not much explicit discussion of tests for relevance within the literature, the reasons we have for being suspicious of some forms of aggregation can be instructive for finding relevant claims. One intuitive thought is that a complaint of lesser strength may be relevant to a complaint of greater strength if aggregating the lesser complaint against the greater one does not give rise to any of the higher-level moral concerns I sketched in §3 of this paper.

15 I have presented Catastrophic Risk to include a low risk short-term intervention. However, everything that I say about how tiny risks relate to large risks of a harm can also be applied to certainties of a harm (if an intervention can deal in certainties.)

Once we divorce *ex-ante* risks from their *ex-post* outcomes, aggregating claims of small risks against large ones looks like it might run straight into the troubling consequences which motivate anti-aggregationism. Let us first consider the concern that some forms of aggregation violates the separateness of persons or presupposes the existence of a fictional entity which will experience the aggregate sum of harm. Now, on an *ex-post* perspective, aggregating the many complaints against the small risks of a harm does not run into this problem; we really do expect that there will be someone – in fact, *many* people – who will feel the sum of the aggregated complaints. Indeed, these would be the people who would die unless we invested in the AI strategy. But, this is not true of the *ex-ante* perspective. We don't expect anyone to experience the aggregate of the *ex-ante* complaints in favour of the AI strategy; that is, we don't expect anyone to experience the aggregate small changes in prospects. As such, it seems that speaking of an aggregate complaint against the change in *ex-ante* prospects would be very much like speaking of the aggregate complaints of inconvenience in Late Train

Now consider the concern that some forms of aggregation are disrespectful to the holder of the larger competing complaint. Likewise, once we are blind to the expected outcomes of many tiny risks, it is tempting to think that considering complaints against miniscule risks of a harm is disrespectful when someone is facing a large risk of that harm – it just does not seem to take seriously the enormity of what is at stake from the person with the largest complaint.

This line of thought is bolstered by considering explanations of relevance currently offered within the literature, the most comprehensive of which, and the account from which many others take inspiration, is Alex Voorhoeve's (Voorhoeve, 2014, 2017; Lazar, 2018; Steuwer, 2021b; Mann, 2021, 2022). Voorhoeve provides an explanation of how claims become irrelevant based on the agent-relative prerogative. It is plausible to think that individuals are, to some extent, permitted to have stronger concern for their own lives and well-being; it

is for this reason that we are permitted not to sacrifice our lives even if doing so would bring about the better outcome. However, such self-concern has limits; it would be, for example, impermissible if an agent failed to a make a trivial sacrifice, such as incurring a slight sore throat or headache, if doing so would save a life. According to Voorhoeve, it is the limits of such agent-relative prerogatives which demark relevance.

In cases in which there are multiple people with competing claims for assistance, each individual's claim has the power to deprive others of assistance. At this point, each agent with a claim could consider whether they are justified in staking that claim, given that doing so could deprive someone else. Now, given the agent-relative prerogative, they are permitted to deprive someone of assistance even if the cost to themselves of forgoing assistance is smaller. However, if the cost to them of not getting assistance is much smaller than the cost to others of not getting assistance, then it seems it would be impermissible to pursue not incurring this cost – the agent relative prerogative simply does not extend that far. As a result, agents are not morally permitted to stake their claims, thereby rendering them irrelevant. You might think this appealing explanation of relevance gives a potential test for relevance; namely, it seems that for a claim to be relevant to another, it has to be permissible for you to stake it. As such, I will make use of what I call:

> **The Sacrifice Test**. A complaint against a harm, x, is relevant to a competing complaint against a greater harm, y, *only if* when given the opportunity to prevent a patient, B, from incurring y at the cost of incurring x themselves, an agent, A, would be permitted not to prevent B incurring y.[16]

16 The sacrifice test is valanced in terms of complaints against harms, but it can also be construed in terms of claims for benefits.

What does Voorhoeve's account, and the sacrifice test, say about cases of risk? We can see by imagining the following situation:

> **Risky Rescue**: During a stroll, Harry sees a child choking on a sweet on the other side of the road. There is no one else around to help, and the child does not seem to be able to dislodge the sweet herself. Harry knows that if he does nothing, the child is likely, though not certain, to asphyxiate. However, the road between them is a fairly dangerous one with poor visibility. Whilst Harry cannot see any vehicles around at present, if he chose to cross the road, he would be incurring the very small risk that a rogue vehicle would appear and fatally hit him.

I take it most would agree it would be wrong of Harry not to help the young child, even though in doing so he would run a tiny risk of death. As such complaints against tiny risks of a harm fail the sacrifice test of being relevant to complaints against large risks of that harm.[17] Following from this, we should think that in cases like Catastrophic Risk, the one-hundred-million individuals' complaint against a tiny risk compare to the ten's complaints against the large risk in the same way complaints against headaches relate to complaints against death – they are just not relevant.

So, it seems that an *ex-ante* partial-aggregationist cannot appeal to the relevance of complaints against different sized risks to avoid the conclusion that when considering prospects, short-term interventions come out on top. Scepticism about aggregation when combined with the *ex-ante* perspective is

17 It has been pointed out to me that individuals might have different intuition in highly abstracted cases, say, one in which you could prevent someone from choking by entering a low-risk death lottery. Whilst I am not sure I share these intuitions, insofar as they do diverge from those elicited in Risky Rescue, I am inclined to think that we should place greater weight on how people approach risky trade-offs in everyday situations, as opposed to how they might in highly abstracted cases. This may especially be the case when we want our theories to have practical application.

sensitive to the fact that long-term interventions, unlike their short-term counterparts, are generally only able to improve the prospects of any future person by a very tiny amount. Long-term interventions simply generate very weak *ex-ante* complaints. As a result, the longtermist who is sceptical about aggregation must turn to the *ex-post* perspective to defend their longtermism.

The necessity of taking up the *ex-post* perspective should not be a comfortable conclusion for the longtermist; *ex-post* nonconsequentialism faces significant criticism, with many pointing to the fact that it can be counter-intuitively risk sensitive and constraining (Ashford, 2003; Fried, 2012; John, 2014; Verweij, 2015; Frick, 2015), and that it faces problems with decomposition (Hare, 2016). This would place the longtermist in an awkward position. However, more importantly for my purpose, I do not believe that even by adopting an *ex-post* analysis can the longtermist defend a good portion of their long-term interventions.

It should be clear that *ex-post* anti-aggregative moral theories are not sensitive, in the way *ex-ante* theories are, to the fact that long-term interventions can only improve the prospects of each individual by a tiny amount. So long as we expect *some* individual to gain a comparable benefit, or avoid a comparable harm, through long-term interventions, then *ex-post* anti-aggregative moral theories will still view long-term interventions as generating strong complaints in their favour.

In fact, an *ex-post* analysis may favour some long-term interventions. Consider, again, Catastrophic Risk. It might be tempting to analyse this case, *ex-post*, as such: if Arthur invests in the risky medical treatment, then in expectation none of the patients will die and one-hundred members of the general population will die; and, if Arthur invests in the AI research, we know ten of the patients will die and, in expectation, fifty members of the general population will die; so, as Arthur expects to generate forty fewer complaints of death, *ex-post*, by investing in the AI research, then he is obligated to choose that option.

Whilst this evaluation is tempting, it is not available to the *ex-post* anti-aggregationist. This manner of evaluating the complaints present in Catastrophic Risk would be appropriate if each of the one-hundred-million had an *independent* risk of incurring the AI related death and suffering – let's call this variant of the case 'Catastrophic Risk (independent)'. If the risk was independent, this would allow us to calculate the expected outcome in which fifty extra people die if we do not invest in the preventative intervention.[18]

However, when it comes to long-term interventions, the risks we are concerned with are not independent. Indeed, this is true of most interventions, and it is especially true of many of the interventions favoured by longtermists, for example those which seek to mitigate global catastrophic risk, s-risks, pandemic risks, or global warming risks. Take the example used in Catastrophic Risk of AI risk – the risk we are exposed to of such an event occurring is actualised in any given individual is connected. If the risk actualises in one person, then it is likely to actualise in many people. The outcome is binary, or at least close to it. Either these events occur, thereby causing many to suffer, or they do not occur and no one suffers for it.

The problem, for the longtermist, is that this complicates the picture of how to calculate what complaints there are *ex-post*. Recall the evaluation sketched above. Putting aside the fact that the numbers in this case are obviously contrived, it is incorrect to say that Arthur *expects* to save forty more lives by investing in the AI research – even if it correct to say that, *in expectation*, he will. This is because it is incorrect to say, *ex-post*, that one-hundred members of the public will die if Arthur fails to attend to the catastrophic AI risk (or that fifty members of the public will die if Arthur does.) There is no possible outcome in which one-hundred (or fifty) people die. Either, with a 99.99995% chance, no

18 In fact, in this case there would be one-hundred-million complaints against death with each discounted as a slightly different rate. However, as in Catastrophic Risk (independent) we have good reason to expect the expectation to occur (namely, fifty deaths), I will simply refer to the fifty complaints of death for the sake of simplicity.

people die, or, with a 0.00005% (that is, a five-in-ten-million) chance, one-hundred-million would die.

If we were permitted to aggregate complaints, the evaluation of this case would be the same as cases in which the risk is not distributed across the potential outcomes in this highly concentrated manner. That is, we could simply work out the *ex-post* complaints if the AI risk did actualise, aggregate them, and then discount the sum of them by the difference made to the improbability of it occurring. So, we would have an aggregate complaint worth one-hundred-million complaints of death, and we would then discount the sum of these complaints by their improbability, to get a complaint roughly worth fifty complaints of death.

However, if we take scepticism about aggregationism seriously, then we cannot apply this sort of thinking. Rather, we must consider all – one-hundred-million – potential complaints of harm individually, and then discount each by the difference the intervention made to its improbability of occurring. So, we are left with one-hundred-million individual complaints of death, each of which is discounted by 99.99995%. On the other hand, the ten patients have very strong *ex-post* complaints: if they do not get the treatment they will die, and if they do they will, almost certainly, survive. So, *ex-post*, each of them have the only ever-so-slightly discounted complaint of death.

The *ex-post* fully non-aggregationist will compare each of the ten slightly discounted complaints of deaths that would occur if Arthur were to choose the AI research option against each of the one-hundred-million massively discounted complaints of death if the medical treatment was chosen. It should be clear that going through this process of pairwise comparison, the non-aggregationist would mandate the treatment option over the AI research option. That is, they would oblige us to engage in the short-term analogue. Like the *ex-*

*ante* theorist, the *ex-post* theorist will view the complaints generated by these types of long-term interventions as uncompetitively weak.

What of the *ex-post* partial-aggregationist? I claim that aggregating the highly discounted complaints of the one-hundred-million against the only slightly discounted claims of the ten will give rise to the sort of morally dubious implication which ground scepticism about aggregation. This is because Arthur, does not *reasonably expect* to save any lives by failing to save the ten.

As the distribution of chances across the outcomes in Catastrophic Risk changes what Arthur can expect to result from his actions, it also changes the sort of justifications he can offer to the ten. Whilst in Catastrophic Risk (independent) Arthur does not disrespect the ten because he can offer a justification grounded in the fact that they expect to save fifty lives, in Catastrophic Risk, Arthur cannot do so. It is not the case that Arthur can turn to the ten and justify his decision not save them from death on the basis that, if he were to help them, then he would expect to bring about a state of affairs in which fifty other people die who could have been saved. Quite the opposite, Arthur is almost certain of the fact that if he were to save the ten, he would not be bringing about a state of affairs in which anyone died when they otherwise would not have. Without such a justification, it is not clear that considering such massively discounted claims is not disrespectful in the context of the ten's significant claims.

Moreover, it is also not the case that we expect a real individual to bear the actual sum of harm which is grounding the aggregated complaint of the one-hundred-million. Again, if Arthur were to fail to attend to the Catastrophic Risk, he would not expect there to be anyone who would die as a result of it. As a result, any talk of aggregate claims in Catastrophic Risk seems similar to attributing the harm to the fictional entity that we did in Late Train.

Turning to the reasons we had for thinking such aggregation was not permitted *ex-ante*, it seems they also suggest aggregation is not permitted in this case. Consider, again, Voorhoeve's account of irrelevance and the sacrifice test. It seems that it would not be permissible for me to forgo saving someone from death (never mind ten people!) because doing so would make it a five-in-ten-millionth more likely to bring about a state of affairs in which I die. As such, it seems that massively discounted complaints of a harm are not relevant to slightly discounted, or non-discounted, complaints of that harms.[19] So both an *ex-post* fully non-aggregative and *ex-post* partially-aggregative moral theory will view the complaints in favour of the long-term analogue in Catastrophic Risk as very weak, obliging us to pick the short-term analogue.

The arguments in this section, thus phrased, show that the typical reasons we provide for thinking that aggregation is impermissible apply to Catastrophic Risk. However, it should be clear that such reasons point to there being a genuine normative difference between cases like Catastrophic Risk and Catastrophic Risk (independent). As such, the arguments of this section can be viewed as rejecting a claim like 'Equal Treatment for Equal Statistical Loss' (Steuwer, 2021b: 119). Here is the general thought: mathematical expectations are not normatively significant in and of themselves. Rather, statistical loss, as informed by considerations of the expected value of an outcome, is a useful metric for what cases ought to be treated alike insofar as it tracks what we expect to occur (and, therefore, the justifications we can offer in an intervention's favour, and so on). As the expectation of an intervention can differ from what we can reasonably expect to occur from it, it should not be surprising that simply

19 Thank you to [redacted] for suggestion the following objection in conversation: as the arguments I present for the irrelevance of the ex-post claims in favour of AI interventions seem to rest on the fact that these claim are highly discounted, one might wonder if the longtermist could get around this conclusion by making the relevance assessment before they discount the complaints. In response, I would point out that such a procedure does not seem available to the ex-post partial-aggregationist; adopting this procedure would cause them to run into familiar problems which discounting was introduced to avoid. Consider again the plane example from footnote 12. If the relevance assessment were made before the complaints were discounted by the difference the intervention made to the likelihood of the harm occurring, then the presence of the undiscounted ex-post complaint against death would render other claims irrelevant..

knowing that, in expectation, two interventions have the same value does not guarantee that we should treat them the same.

To pull together the loose threads of this discussion, I have argued that *ex-ante* anti-aggregative moral theories will choose short-term interventions over long-term interventions given that short-term intervention create significantly bigger changes in the prospects of individuals than long-term interventions. *En route* to this conclusion, I have also claimed that, *ex-ante*, complaints against small risks of a harm are not relevant to complaints against large risks of that, or a comparable, harm.

I have then argued that whilst *ex-post* anti-aggregative moral theories can accommodate a preference for some long-term interventions, they are unable to do so for a key class of long-term interventions. Due to the distribution of risk across the outcomes featured in a number of interventions, including those which seek to manage catastrophic risk, the complaints of future people who might be benefitted by such interventions are massively discounted. As such *ex-post* anti-aggregative moral theories will prefer short-term interventions to those which seek to mitigate catastrophic risk.

Critically, *ex-post* anti-aggregative moral theories *can* display a preference for some long-term interventions, but this is only when the intervention in question is such that you *reasonably expect* to prevent a comparable harm by engaging in it. Otherwise, to aggregate the *ex-post* complaints in favour of such an intervention would plausibly violate the separateness of persons and disrespect those with greater claims to our assistance. For those sceptically-minded longtermists, the important question becomes an empirical one: are there any very far-future interventions which are such that we can reasonably expect to save a life with them?

## 5. Tug of war and conclusions

The argument presented in this paper pulls in two distinct directions. For those unwilling to give up their scepticism about aggregation, this paper's conclusion might be that such scepticism should extend to longtermism and long-term interventions. It seems that, on both an *ex-ante* and *ex-post* reading, it is hard to value the way in which long-term interventions bestow goods without allowing for aggregation. Certainly, *ex-ante* anti-aggregationist moral theories seem to systematically prefer the goods bestowed by short-term interventions. And, even if we were to adopt an *ex-post* perspective, it seems that many of the sorts of interventions which longtermists are most concerned with – those, for example, which seek to mitigate global catastrophic risk – are excluded.

But, of course, one might also see the conclusions of this argument as contributing to the growing literature raising suspicion against both non-aggregative and partially-aggregative moral theories (Norcross, 1997, 1998; Reibetanz, 1998; Dougherty, 2013; Tomlin, 2017; Horton, 2017, 2018, 2020). By failing to value the goods longtermist interventions can bestow, one might think that this is further evidence of anti-aggregationism being insufficiently sensitive to axiological stakes. Likewise, by excluding those cases analogous to Catastrophic Risk – which very plausibly would include many present day practices, including those which mitigate natural disaster risk – this may be simply more evidence pointing to the fact that anti-aggregative moral theorising is deeply unpracticable. From such a perspective, this paper is not really about longtermism nor the future. It's really a paper about aggregation, wrapped up in the useful counterexample of long-term interventions.

At this point, it is easy to think that we are confronted with a decision about which of our philosophical commitments would be easiest to give up. Would it be easier to throw to the side our intuitions in Late Train, alongside the separateness of persons? Or, perhaps, it may be easier to give up our feelings of

obligation those in the far-future? Perhaps this paper suggests a third, less obvious conclusion. The realisation that these various intuitions and commitments are in conflict does not make any of them less attractive. It still seems to me that a plausible moral theory ought not compel us to let the man die in Late Train. It also still seems to me that a plausible moral theory ought to, at least, permit us to attend to risks like catastrophic risk, s-risks, or natural disaster risks. This paper can be viewed as suggesting, however, that there might not be a moral theory that can consistently do both.

## 6. References


Adams, Fred C. (2008). Long-Term Astrophysical Processes, in Nick Bostrom and Milan Cirkovic (eds). *Global Catastrophic Risks*, Oxford: Oxford University Press

Ashford, Elizabeth. (2003). The Demandingness of Scanlon's Contractualism, *Ethics*, 113(2): 273-302

Balfour, Dylan. (2020). Pascal's Mugger Strikes Again, *Utilitas*, 33(1): 118-24

Beckstead, Nick. (2013). *On the overwhelming importance of shaping the far future*, PhD thesis. Department of Philosophy, Rutgers University, doi:10.7282/T35M649T

———— (2019). A Brief Argument for the Overwhelming Importance of Shaping the Far Future, in Hilary Greaves and Theron Pummer (eds). *Effective Altruism: Philosophical Issues*, Oxford: Oxford University Press: 80-98

Beckstead, Nick and Teruji Thomas. (2021). *A paradox for tiny probabilities and enormous values*, GPI Working Paper, retrieved from https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/

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

———— (2009). Pascal's Mugging, *Analysis*, 69(3): 443-445

———— (2013). Existential Risk Prevention as Global Priority, *Global Policy*, 4: 15-31

Carlson, Erik. (2000). Aggregating Harms – Should We Kill to Avoid Headaches? *Theoria*, 66(3): 246-255

Daniels, Norman. (2015). Can There Be Moral Force to Favoring an Identified over a Statistical Life? In in I. Glenn Cohen, Norman Daniels, and Nir Eyal (eds). *Identified versus Statistical Lives: An Interdisciplinary Perspective*, Oxford: Oxford University Press: 110-123

Dorsey, Dale. (2009). Headaches, Lives and Value, *Utilitas*, 21(1): 36-58

Dougherty, Tom. (2013). Aggregation, Beneficence, and Chance, *Journal of Ethics and Social Philosophy*, 7(2): 1-19

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

Fried, Barbara H. (2012). Can Contractualism Save Us from Aggregation? *Journal of Ethics*, 16(1): 39-66

van Gils, Aart and Patrick Tomlin. (2020). Relevance Rides Again? Aggregation and Local Relevance, in David Sobel, Peter Vallentyne and Steven Wall (eds). *Oxford Studies in Political Philosophy*, 6, Oxford: Oxford University Press: 221-256

Greaves, Hilary, and Will MacAskill. (2021). *The case for strong longtermism*, GPI Working Paper, retrieved from https://globalprioritiesinstitute.org/wp-content/uploads/The-Case-for-Strong-Longtermism-GPI-Working-Paper-June-2021-2-2.pdf

Hare, Caspar. (2016). Should We Wish Well to All?*, The Philosophical Review*, 124(4): 451-472

Horton, Joe. (2017). The All or Nothing Problem, *Journal of Philosophy*, 114(2): 94-104

———— Aggregation, Complaints, and Risk, *Philosophy & Public Affairs*, 45(1): 54-81

———— (2018). Always Aggregate, *Philosophy & Public Affairs*, 46(2): 160-174

———— (2020). Aggregation, Risk, and Reductio, *Ethics*, 130(4): 514-529

———— (2021). Partial aggregation in ethics, *Philosophy Compass*, doi:10.1111/phc3.12719

John, Stephen. (2014). Risk, Contractualism and Rose's "Prevention Paradox", *Social Theory and Practice*, 40(1): 28-50

Kamm, F. M. (1993). *Morality, Mortality*, 1, New York: Oxford University Press

Kumar, Rahul. (2003). Reasonable reasons in contractualist moral argument, *Ethics*, 114(1): 6-37

Lazar, Seth. (2018). Limited Aggregation and Risk, *Philosophy & Public Affairs*, 46(2): 117-159

Lazar, Seth and Chad Lee-Stronach. (2019). Axiological Absolutism and Risk, *Noûs*, 53(1): 97-113

Mann, Kirsten. (2021). The Relevance View: Defended and Extended, *Utilitas*, 33(1): 101-110
———— (2022). Relevance and Nonbinary Choices, *Ethics*, 132(2): 382-413

McMahan, Jeff. (2018). Doing Good and Doing the Best, in Paul Woodruff (ed). *The Ethics of Giving: Philosophers' Perspectives on Philanthropy*, New York: Oxford University Press: ch.3

Mogensen, A.L. (2020). Moral demands and the far future, *Philosophy and Phenomenological Research*, 00:1-19, doi:10.1111/phpr.12729

Norcross, Alastair. (1997). Comparing Harms: Headaches and Human Lives, *Philosophy & Public Affairs*, 26(2): 135-67

———— (1998). Great Harm From Small Benefits Grow: How Death Can Be Outweighed by Headaches, *Analysis*, 58(2): 152-158

Nozick, Robert. (1974). *Anarchy, State, and Utopia*, New York: Basic Books

Ord, Toby. (2020). *The Precipice: Existential risk and the future of humanity*, London: Bloomsbury

Otsuka, Michael. (2006). Saving Lives, Moral Theory, and the Claims of Individuals, *Philosophy & Public Affairs*, 34(2): 109-135

———— (2015). Risking life and limb, in I. Glenn Cohen, Norman Daniels, and Nir Eyal (eds). *Identified versus Statistical Lives: An Interdisciplinary Perspective*, Oxford: Oxford University Press: 77-93

Pummer, Theron. (2016). Whether and Where to Give, *Philosophy & Public Affairs*, 44(1): 77-95

Rawls, John. (1971). *A Theory of Justice*, Cambridge, MA: Belknap Press of Harvard University Press

Reibetanz, Sophia. (1998). Contractualism and Aggregation, *Ethics*, 108(2): 296-311

Rüger, Korbinian. (2020). Aggregation with Constraints, *Utilitas*, 32(4): 454-471

Scanlon, T. M. (1998). *What We Owe to Each Other*, Cambridge, MA: Belknap Press of Harvard University Press

Steuwer, Bastian. (2020). *One-by-one: moral theory for separate persons*. PhD thesis, Department of Philosophy, Logic and Scientific Method, The London School of Economics and Political Science, retrieved from http://etheses.lse.ac.uk/4149/1/Steuwer__One-by-one-moral-theory.pdf

———— (2021a). Aggregation, Balancing, and Respect for the Claims of Individuals, *Utilitas*, 33(1): 17-34

———— (2021b). Contractualism, Complaints, and Risk, *Journal of Ethics and Social Philosophy*, 19(2): 111-147

Tadros, Victor. (2019). Localised restricted aggregation, *Oxford Studies in Political Philosophy*, 5: 171-204

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

Temkin, Larry S. (2012) *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, New York: Oxford University Press

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

Verweij, Marcel. (2015). How (Not) to Argue for the Rule of Rescue: Claims of Individuals Versus Group Solidarity, in I. Glenn Cohen, Norman Daniels, and Nir Eyal (eds). *Identified versus Statistical Lives: An Interdisciplinary Perspective*, Oxford: Oxford University Press: 137-149

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

———— (2017). Why One Should Count Only Claims with Which One Can Sympathize, *Public Health Ethics*, 10(2): 148-156

Wilkinson, Hayden. (2022). In defence of fanaticism, *Ethics*, 132(2): 445-77