

ORIGINAL ARTICLE

Longtermism and aggregation

Emma Curran

University of Oxford, Oxford, UK

CorrespondenceEmma Curran, University of Oxford,
Oxford, UK.Email: emma.curran@philosophy.ox.ac.uk**Funding information**Arts and Humanities Research Council,
Grant/Award Number: AH/R012709/1**Abstract**

Advocates of longtermism point out that interventions which focus on improving the prospects of people in the very far future will, in expectation, bring about an astronomical amount of good (or agent-neutral value). As such, longtermists claim we have compelling moral reason to engage in long-term interventions. In this paper, I show that longtermism is in conflict with plausible deontic scepticism about aggregation. I do so by demonstrating that, from both the *ex-ante* and *ex-post* perspectives, longtermist interventions generate extremely weak claims of assistance from future people.

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.¹ Advocates of *longtermism* point out that, in expectation, long-term interventions bring about far more good, or agent-neutral value, than short-term interventions, which tend to those living now or soon.

This claim is 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 colonising other planets, which we could inhabit for up to 100 trillion years (Adams, 2008, p. 39; Beckstead, 2019, p. 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

¹ Particularly significant discussion of longtermism can be found from Bostrom (2003, 2013), Beckstead (2013), Ord (2020), Andreas Mogensen (2021), and Greaves and MacAskill (forthcoming).

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2025 The Author(s). *Philosophy and Phenomenological Research* published by Wiley Periodicals LLC on behalf of Philosophy and Phenomenological Research LLC.

more reasonable estimate would be at least one quadrillion ([forthcoming](#), p. 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 the patients have exhausted all other treatment options. If they do not have this procedure, they will all certainly 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. Even though by intervening on the large group, Arthur would only improve the expected wellbeing of each individual by a small amount, this would, in expectation, do a lot of good on the group level. 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. Firstly, 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, secondly, even the slightest increase to the prospects of each future person generates a vast amount of good, in expectation. As such, long-term interventions which improve the prospects of each future person 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.² Such a conclusion, longtermists argue, provide us with compelling, if not decisive, reasons of beneficence to choose such interventions. Indeed, given the enormity of

² Greaves and MacAskill call this claim, or something similar to it, “axiological strong longtermism” ([forthcoming](#), p. 6). Axiological strong longtermism has, however, been questioned, spawning a literature on expected utility theory and

the axiological stakes involved, some philosophers argue we have a moral obligation to invest in long-term interventions.³

My aim in writing this paper is to highlight a conflict between longtermism and, what I think is, plausible deontic scepticism about aggregation. I show that the beneficence argument for longtermism relies on an assumption about the permissibility of aggregation; deontic scepticism about aggregation not only undermines an obligation to invest in long-term interventions, but also the permissibility of doing so.

Two clarifications need to be made. First, the considerations I present in this paper are targeted at arguments for longtermism grounded in beneficence. They do not speak against any prudential reasons we might have to engage in long-term interventions (Gustafsson and Kosonen, [forthcoming](#); Scheffler, 2018). Second, this paper also ignores considerations of non-identity. In doing so, it shows that our obligations to future people are not wholly determined by the fate of the non-identity problem. We can assume that future people have straightforward claims to our assistance, just as those alive do, and yet have good reason to forgo helping them.

In §2, I outline concerns about aggregation and the *ex-ante/ex-post* distinction, respectively. Once introduced, I proceed, in §3 and 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. In §5, I respond to potential criticism of my argument, before discussing its upshots in §6.

2 | AGGREGATION AND RISK

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 a remote break for the train. She can either pull the lever, causing all the passengers, of which there are N , to be somewhat late, or allow the train to fatally run over the man.⁴

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.⁵

fanaticism, see Bostrom (2009), Balfour (2020), Beckstead and Thomas (2023), and Wilkinson (2022). David Thorstad has also convincingly argued that the expected value of mitigating existential risk is far lower than longtermists might claim (2023).

³ See, in particular, Greaves and MacAskill ([forthcoming](#), p. 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.

⁴ Another memorable example of this type is Scanlon's Transmitter Room (1998, p. 235).

⁵ 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.

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 be best to allow the man to be run 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.⁶ 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.⁷ Nonetheless, I believe much of what I say about the conflict between deontic anti-aggregationism and longtermism will apply, *mutatis mutandis*, to axiological anti-aggregationism.

The second diagnosis allows that for some N it might 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.⁸

Scepticism about aggregation also comes from other sources, for example higher-level moral principles like the *separateness of persons*.⁹ 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, pp. 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, pp. 144–163).¹⁰

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.

⁶ For proposals and discussion of axiological anti-aggregative moral theories, see Carlson (2000), Dorsey (2009), Temkin (2012), and Lazar and Lee-Stronach (2019).

⁷ I, in part, place axiological accounts to the side due to their apparent implausibility. See Horton (2021) for an overview of their problems.

⁸ Most famously contractualist follow the '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, 2003a). Contractualism is, however, compatible with axiological scepticism.

⁹ Concern about the separateness of persons and aggregating across persons comes from Rawls (1971, pp. 26–27).

¹⁰ 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.¹¹ 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 competing complaint through a process of pairwise comparison. 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.¹²

Partially-aggregative moral theories are similar to their non-aggregative cousins, except they permit aggregation in some instances.¹³ Permitting aggregation of some complaints, whilst generally prohibiting it, has the benefit of both respecting our convictions in cases like Late Train, whilst also cohering with other ground-level moral intuitions. 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. In such a case, we would say that a complaint of quadriplegia is *relevant* to a complaint of death. 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. Generally, partially-aggregative accounts tell us to choose the option which satisfies the greatest sum of strength-weighted, relevant complaints.¹⁴

The question of this paper is what such anti-aggregative theories of morality have to say about longtermism. Before I can answer this question, however, I must introduce one final distinction – one concerning risk.

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 *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 whose claims we attend to. When we assess decisions *ex-ante*, we are considering the claims of

¹¹ Henceforth, I will simply talk of complaints.

¹² Contractualism is one example of such a theory (Scanlon, 1998).

¹³ For proposals, and discussion, of such views see: Voorhoeve (2014, 2017), Lazar (2018), Tadros (2019), R ger (2020), Steuwer (2021a), and Mann (2021). 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).

¹⁴ The initial proponent of this general type of view is 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.

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 individuals *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 implemented 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. 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, p. 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 *someone* will incur a harm, not that any particular person will (Otsuka, 2015, pp. 86–89). As such, we discount each of these complaints by the difference the intervention would make to their improbability of occurring.

With the distinction between *ex-ante* and *ex-post* moral theories at hand, we can now outline the four types of anti-aggregative 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.

3 | THE EX-ANTE ASSESSMENT

I can now demonstrate that the argument that beneficence requires choosing those interventions which most improve the prospects of future people relies on full blown deontic aggregation. 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 the patients have exhausted all other treatment options. If they do not have this procedure, they will all certainly 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-aggregative moral theories will favour treating the ten – in fact, they would claim that it was *obligatory* to invest in the treatment option.

The result that the *ex-ante* fully non-aggregative would not be persuaded by long-term interventions should not be surprising; philosophers have noted for some time that *ex-ante* non-aggregative theories have this property (Frick, 2015). However, what of the *ex-ante* partial-aggregative? They will come to the same conclusion unless they believe that complaints against small risks of a harm are relevant to complaints against much larger risks of that harm, such that, *ex-ante*, we can permissibly aggregate them against one another.¹⁵ I find this 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 §2 of this paper.

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 violate the separateness of persons or presupposes the existence of a fictional entity which will experience the aggregate sum of harm. Now, from 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 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 (2014, 2017). 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 make a trivial sacrifice, such as incurring a slight sore

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.)

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. This appealing explanation of relevance gives a potential test for relevance, which I will call:

The Sacrifice Test. A complaint against *x* is relevant to a competing complaint against *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

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' complaints 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. The *ex-ante* partial-aggregationist must also prefer short-term interventions.

Scepticism about aggregation when combined with the *ex-ante* perspective is 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

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

¹⁷ It has been pointed out to me that individuals might have different intuitions 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.

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.

4 | THE EX-POST ASSESSMENT

The necessity of taking up the *ex-post* perspective should not be a comfortable conclusion for the longtermist; *ex-post* nonconsequentialism violates the extremely plausible *ex-ante* Pareto principle, and it has been shown to be counter-intuitively risk sensitive and constraining (Ashford, 2003; Fried, 2012; John, 2014; Verweij, 2015; Frick, 2015), and it faces problems with decomposition (Hare, 2016). This would place the longtermist in an awkward position. However, more importantly – and more surprisingly – I do not believe that even by adopting an *ex-post* analysis can longtermism be fully defended.

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 claims 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. Again, this line of thought should not be surprising; it has been acknowledged before (Frick, 2015).

Whilst this evaluation is tempting, it is not available to the *ex-post* anti-aggregationist in the case of long-term interventions. 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

¹⁸ In fact, in this case there would be one-hundred-million complaints against death with each discounted at a slightly different rate, following a binomial distribution. 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.

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

If we take scepticism about aggregationism seriously, then we must consider all one-hundred-million potential complaints of harm individually – after all this is the number of people that would be killed in the worst-case scenario. We would 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 implications 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 to 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. As such, 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 harm. 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.

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 creates 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 numerous 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?

One might be initially surprised that the *ex-post* perspective assesses cases like Catastrophic Risk and Catastrophic Risk (independent) so differently given they share the same expected value. Yet, once properly considered, I think this difference is not as counterintuitive as it might at first seem. Mathematical expectations are not normatively significant in and of themselves. Rather, statistical loss 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 for our actions. 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 knowing that, in expectation, two interventions have the same value does not guarantee that we should treat them the same.

5 | OBJECTIONS AND RESPONSES

In this section of the paper, I want to address two objections which I suspect one might raise against the argument presented – and, indeed, some objections that have actually been raised whilst presenting this paper. The first is that there is an *important disanalogy* between Catastrophic Risk and real long-term interventions, which limits the scope of my argument. This disanalogy is that the risk which is being mitigated in Catastrophic Risk is of an antecedently very low level, whilst many of the sort of risks which longtermists are interested in mitigating – such as asteroid impacts, as so forth – have plausibly high antecedent levels. The concern, then, is that the argument presented only shows an inconsistency between anti-aggregative moral theories and long-term interventions which aim to mitigate already small risks.

This objection is mistaken. The reason that both the *ex-ante* and *ex-post* complaints are so weak is not primarily because the risks long-term interventions are mitigating are tiny to begin with, but rather that they can only reduce these risks by a tiny amount. Now, yes, that a risk is antecedently low is going to give rise to this. But this is not the main reason why long-term interventions are associated with small reductions in risk, rather it is because of our limited ability to predictably control the very far future. Regardless of whether the risk we are looking to mitigate is antecedently high or low, we cannot do anything about it with much certainty. As such, the argument follows irrespective of the antecedent risk level.

There are two small caveats that must be noted. First, whilst it is a general rule that we cannot control which states-of-affairs will obtain in the future with much certainty, it is not a rule without exceptions. Some long-term interventions, such as leaving money in trust, might allow us to impact the future with high probability. Such interventions are not vulnerable to my argument. Second, in cases in which the risk facing the future are antecedently large, my argument might be weakened by considerations of priority-weighted complaints. In such a case, the fact that future people are expected to be badly off might mean that we should assign greater weight to improving their wellbeing (even if only by a small amount, in expectation). Such concerns would need to be settled elsewhere.

The second objection is that there is an *alternative way of discounting claims ex-post* which avoids this dim view of long-term interventions. The method of assessing *ex-post* claims I use involves discounting the *ex-post* claims by the difference the intervention made to the probability of the harm occurring, then – in the case of the partial-aggregationist – aggregating relevant discounted complaints. It is because we discounted first that the many *ex-post* complaints were deemed irrelevant to the competing claims in favour of the short-term analogue. Therefore, an *ex-post* partially aggregative justification of longtermism could be found if we instead made the relevance judgements first, then discounted the *ex-post* claims.¹⁹ The many complaints of death from AI risk would then be relevant to the competing complaints, and could be aggregated together and then discounted, in turn outweighing the claims in favour of the short-term analogue.

In response, I would point out that adopting such a procedure would come at a high cost which may sacrifice the plausibility of partial aggregation. Consider the example of a policy introducing a new flight path over a remote island, which would reduce the time of many flights dramatically. However, allowing planes to fly over this island will impose a tiny risk of death-by-falling-plane on the single person inhabiting the island.²⁰ It seems that such a policy would be justifiable given the convenience of the new flight path. However, if we made the relevance judgements about the *ex-post* claims of convenience and death *before* discounting the claims of death, then the claims of convenience in favour of flight path will be deemed irrelevant. As such, such a theory will tell us we could never impose a risk of a serious harm to bestow less significant benefits – regardless of how tiny that risk of harm is and the fact that in expectation no one will face the harm. This strains credulity and would make *ex-post* anti-aggregative moral theories unworkably risk averse.

¹⁹ Thank you to Kacper Kowalczyk for bringing this problem to my attention.

²⁰ Examples of this nature are discussed extensively in the literature and is the reason Otsuka (2015) initially proposes discounting *ex-post* claims. Here I show that even if we discount claims, if we do not do so before we make relevance comparisons, then we will run into the same problem.

6 | 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 – at least in terms of beneficence.

This conclusion might seem especially significant given that throughout this paper I have placed non-identity concerns to one side.²¹ It seems to me that it has often been assumed that for complaints based theorists, the question of the extent of our obligations to those in the future is determined in large part by our answer to the non-identity problem. That is, if one assumes that far-future people straightforwardly generate the sort of claims or complaints that currently existing people do, then we have considerable, potentially decisive, moral reason to help them. Yet, the argument of this paper demonstrates that to a large extent our answer to the non-identity problem is unimportant for those who are sceptical about aggregation. Even if future people generate complaints, they will be, in a vast array of cases, extremely weak. This is not anything to do with their moral status as 'merely possible people', it is simply a contingent result of the long-term interventions available to us, and the uncertainty with which we can shape the far-future.

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.²² As is often said, one man's *modus ponens* is another man's *modus tollens*. 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.²³ 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 to 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.²⁴

²¹ See Parfit (1982, 1987, Ch.16) for an outline of the non-identity problem, alongside Kumar (2003b) and Hare (2007) for complaint-based responses.

²² Criticisms can be found from Norcross (1997, 1998), Reibetanz (1998), Dougherty (2013), Tomlin (2017), and Horton (2017, 2018, 2020).

²³ Whilst a discussion of this is beyond the remit of this paper, I address this objection, alongside the practicability of contractualist approaches to longtermism, in Curran (forthcoming).

²⁴ A special thanks must be given to Arif Ahmed who has endured reading countless drafts of this paper whilst never failing to give thoughtful feedback. I am also grateful to Stephen John, Theron Pummer, Richard Holton, Hilary Greaves,

REFERENCES

- Adams, F. (2008). Long-term astrophysical processes. In N. Bostrom, & M. Cirkovic (Eds.), *Global catastrophic risks* (pp. 33–47). Oxford University Press.
- Ashford, E. (2003). The demandingness of Scanlon's contractualism. *Ethics*, *113*, 273–302. <https://doi.org/10.1086/342853>
- Balfour, D. (2020). Pascal's mugger strikes again. *Utilitas*, *33*, 118–124. <https://doi.org/10.1017/S0953820820000357>
- Beckstead, N. (2013). *On the overwhelming importance of shaping the far future*, PhD thesis. Department of Philosophy, Rutgers University. <https://doi.org/10.7282/T35M649T>
- Beckstead, N. (2019). A brief argument for the overwhelming importance of shaping the far future. In H. Greaves & T. Pummer (Eds.), *Effective altruism: Philosophical issues* (pp. 80–98). Oxford University Press.
- Beckstead, N. and Thomas, T. (2023). A Paradox for tiny probabilities and enormous values. *Nous*, *58*, 431–455. <https://doi.org/10.1111/nous.12462>
- Bostrom, N. (2003). Astronomical waste: The opportunity cost of delayed technological development. *Utilitas*, *15*, 308–314. <https://doi.org/10.1017/S0953820800004076>
- Bostrom, N. (2009). Pascal's mugging. *Analysis*, *69*, 443–445. <https://doi.org/10.1093/analys/69/4>
- Bostrom, N. (2013). Existential risk prevention as global priority. *Global Policy*, *4*, 15–31. <https://doi.org/10.1111/1758-5899.12002>
- Carlson, E. (2000). Aggregating harms – Should we kill to avoid headaches? *Theoria*, *66*, 246–255. <https://doi.org/10.1111/j.1755-2567.2000.tb01167.x>
- Curran, E. (forthcoming). Longtermism and the complaints of future people. In H. Greaves, J. Barrett, & D. Thorstad (Eds.), *Essays on longtermism*. Oxford University Press.
- Dorsey, D. (2009). Headaches, lives and value. *Utilitas*, *21*, 36–58. <https://doi.org/10.1017/S095382080800335X>
- Dougherty, T. (2013). Aggregation, beneficence, and chance. *Journal of Ethics and Social Philosophy*, *7*(2), 1–19. <https://doi.org/10.26556/jesp.v7i2.73>
- Frick, J. (2015). Contractualism and social risk. *Philosophy & Public Affairs*, *43*, 175–223. <https://doi.org/10.1111/papa.12058>
- Fried, B. (2012). Can contractualism save us from aggregation? *Journal of Ethics*, *16*, 39–66. <https://doi.org/10.1007/s10892-011-9113-3>
- van Gils, A. & Tomlin, P. (2020). Relevance rides again? aggregation and local relevance. In D. Sobel, P. Vallentyne and S. Wall (Eds.), *Oxford Studies in Political Philosophy* (Vol.6, pp. 221–256). Oxford University Press.
- Greaves, H. & MacAskill, W. (forthcoming). The case for strong longtermism. In H. Greaves, J. Barrett, & D. Thorstad (Eds.), *Essays on longtermism*. Oxford University Press.
- Gustafsson, J. & Kosonen, P. (forthcoming). Prudential longtermism. In H. Greaves, J. Barrett, & D. Thorstad (Eds.), *Essays on longtermism*. Oxford University Press.
- Hare, C. (2007). Voices from another world: Must we respect the interests of people who do not, and will never, exist? *Ethics*, *117*, 498–523. <https://doi.org/10.1086/512172>
- Hare, C. (2016). Should we wish well to all? *The Philosophical Review*, *124*, 451–472. <https://doi.org/10.1215/00318108-3624764>
- Horton, J. (2017). Aggregation, complaints, and risk. *Philosophy & Public Affairs*, *45*, 54–81. <https://doi.org/10.1111/papa.12084>
- Horton, J. (2018). Always aggregate, *Philosophy & Public Affairs*, *46*, 160–174. <https://doi.org/10.1111/papa.12116>
- Horton, J. (2020). Aggregation, risk, and reductio. *Ethics*, *130*, 514–529. <https://doi.org/10.1086/708534>
- Horton, J. (2021). Partial aggregation in ethics. *Philosophy Compass*. <https://doi.org/10.1111/phc3.12719>
- John, S. (2014). Risk, contractualism and rose's "Prevention Paradox". *Social Theory and Practice*, *40*, 28–50. <https://doi.org/10.5840/soctheorpract20144012>
- Kamm, F. M. (1993). *Morality, mortality*, 1. Oxford University Press.
- Kumar, R. (2003a). Reasonable reasons in contractualist moral argument. *Ethics*, *114*, 6–37. <https://doi.org/10.1086/378034>

Andreas Mogensen, Kyle van Oosterum, Owen Clifton, David Thorstad, Tomi Francis, and an anonymous reviewer for this journal for their invaluable comments. I would like to thank audiences at the Global Priorities Institute, the University of Cambridge, the University of Sheffield, the University of Minho, and the PPE Society for their comments.

- Kumar, R. (2003b). Who can be wronged? *Philosophy & Public Affairs*, 3, 99–118. <https://doi.org/10.1111/j.1088-4963.2003.00099.x>
- Lazar, S. (2018). Limited aggregation and risk. *Philosophy & Public Affairs*, 46, 117–159. <https://doi.org/10.1111/papa.12115>
- Lazar, S & Lee-Stronach, C. (2019). Axiological absolutism and risk. *Noûs*, 53: 97–113. <https://doi.org/10.1111/nous.12210>
- Mann, K. (2021). The relevance view: Defended and extended. *Utilitas*, 33, 101–110. <https://doi.org/10.1017/S095382082000028X>
- Mann, K. (2022). Relevance and nonbinary choices. *Ethics*, 132, 382–413. <https://doi.org/10.1086/716873>
- Mogensen, A. (2021). Moral demands and the far future. *Philosophy and Phenomenological Research*, 103, 567–585. <https://doi.org/10.1111/phpr.12729>
- Norcross, A. (1997). Comparing harms: Headaches and human lives. *Philosophy & Public Affairs*, 26, 135–167. <https://doi.org/10.1111/j.1088-4963.1997.tb00079.x>
- Norcross, A. (1998). Great harm from small benefits grow: How death can be outweighed by headaches. *Analysis*, 58, 152–158. <https://doi.org/10.1111/1467-8284.00116>
- Nozick, R. (1974). *Anarchy, state, and utopia*. Basic Books.
- Ord, T. (2020). *The precipice: Existential risk and the future of humanity*. Bloomsbury.
- Otsuka, M. (2015). Risking life and limb. In I. G. Cohen, N. Daniels & N. Eyal (Eds.), *Identified versus statistical lives: An interdisciplinary perspective* (pp. 77–93). Oxford University Press.
- Parfit, D. (1982). Future generations: Future problems. *Philosophy & Public Affairs*, 11, 113–72
- Parfit, D. (1987). *Reasons and persons*. Clarendon Press.
- Rawls, J. (1971). *A theory of justice*. Belknap Press of Harvard University Press.
- Reibetanz, S. (1998). Contractualism and aggregation. *Ethics*, 108, 296–311. <https://doi.org/10.1086/233806>
- Rüger, K. (2020). Aggregation with constraints. *Utilitas*, 32, 454–471. <https://doi.org/10.1017/S095382082000014X>
- Scanlon, T. M. (1998). *What we owe to each other*. Belknap Press of Harvard University Press.
- Scheffler, S. (2018). *Why worry about future generations?* Oxford University Press.
- Steuer, B. (2021a). Aggregation, balancing, and respect for the claims of individuals. *Utilitas*, 33, 17–34. <https://doi.org/10.1017/S0953820820000217>
- Steuer, B. (2021b). Contractualism, complaints, and risk. *Journal of Ethics and Social Philosophy*, 19, 111–47. <https://doi.org/10.26556/jesp.v19i2.1024>
- Tadros, V. (2019). Localised restricted aggregation. In D. Sobel, P. Vallentyne & S. Wall (Eds.), *Oxford studies in political philosophy* (Vol. 5, pp. 171–204). Oxford University Press.
- Temkin, L. (2012) *Rethinking the good: Moral ideals and the nature of practical reasoning*. Oxford University Press.
- Thorstad, D. (2023) High risk, low reward: A challenge to the astronomical value of existential risk mitigation. *Philosophy & Public Affairs*, 51, 373–412. <https://doi.org/10.1111/papa.12248>
- Tomlin, P. (2017). On limited aggregation. *Philosophy & Public Affairs*, 45, 232–260. <https://doi.org/10.1111/papa.12097>
- Verweij, M. (2015). How (Not) to argue for the rule of rescue: Claims of individuals versus group solidarity. In I.G. Cohen, N. Daniels, and N. Eyal (Eds.), *Identified versus statistical lives: An interdisciplinary perspective* (pp. 137–149). Oxford University Press.
- Voorhoeve, A. (2014). How should we aggregate competing claims? *Ethics*, 125, 64–87. <https://doi.org/10.1086/677022>
- Voorhoeve, A. (2017). Why one should count only claims with which one can sympathize. *Public Health Ethics*, 10, 148–156. <https://doi.org/10.1093/phe/phw006>
- Wilkinson, H. (2022). In defence of fanaticism. *Ethics*, 132, 445–477. <https://doi.org/10.1086/716869>

How to cite this article: Curran, E. (2025) Longtermism and aggregation. *Philosophy and Phenomenological Research*, 110, 1137–1151. <https://doi.org/10.1111/phpr.70007>