(To appear in *Essays on Longtermism*, Jacob Barrett, Hilary Greaves & David Thorstad (eds), Oxford University Press.)

Shaping humanity's longterm trajectory

Toby Ord[1]

This chapter presents a mathematical framework for classifying and comparing the different kinds of effects present-day actions could have upon the longterm future of humanity. The starting point is the longterm trajectory of humanity, understood as how the instantaneous value of humanity unfolds over time. In this framework, the value of our future is equal to the area under this curve and the value of altering our trajectory is equal to the area between the original curve and the altered curve. This allows us to compare the value of reducing existential risk to other ways our actions might improve the longterm future, such as improving the values that guide humanity, or advancing progress.

Keywords: longtermism, longterm trajectory, existential risk,
future of humanity, advancement, flow utility.

Introduction

Humanity may have a very long time ahead. Our species has already survived about 300,000 years, and the typical species lifespan is roughly one million years (Galway-Witham and Stringer 2018; Barnosky 2011). Some species, such as the horseshoe crab and the nautilus, have survived unaltered for hundreds of millions of years. One kind of upper bound on humanity's longevity comes from the Earth itself, which will remain habitable for roughly one billion more years. But even that is not a hard limit: if our descendants migrate to other stellar systems with newer (or longer lived) suns, they could survive for at least a thousand times as long again (Adams and Laughlin 1997 & 1999). So if we play our cards right, humanity could have a flourishing civilisation that lasts for a time that is almost beyond comprehension.

Longtermism is a moral outlook that takes this vast future seriously (MacAskill 2022, Greaves & MacAskill 2021, Ord 2020: 43–49). It considers how the possibility of making lasting alterations to humanity's longterm future might inform the actions we should take now.

[1] I'd like to thank Nick Beckstead for early discussions about quantifying the value of advancements, and Owen Cotton-Barratt, Will MacAskill, and Finlay Moorhouse for many further conversations that helped this framework come together. This work was funded by a grant from Open Philanthropy.

One clear way humanity's longterm future might be altered is via *existential risks* — risks of irrevocably destroying humanity's longterm potential, such as through extinction or an unrecoverable collapse of civilisation (see Bostrom 2013, and Ord 2020). Examples include the risks of asteroid impacts, supervolcanic eruptions, nuclear war, engineered pandemics, and climate change. Existential catastrophes provide a clear example of how our actions today could have effects that don't simply wash out over time. This is because:

(1) they could occur in our own time,
(2) if they did occur, they rule out the chance of a subsequent recovery, yet
(3) they could be prevented by concerted human action.

This means that an action to increase or decrease existential risk could be targeted at a near-future event, yet have an expected value that scales in proportion to the entire size of humanity's longterm potential.

But avoiding existential risk isn't the only way our actions might shape the longterm future. For example, we may be able to temporarily — or even permanently — speed up progress: allowing people earlier access to the prosperity, technology, or social progress that would have come in later years. And if there is a chance that humanity's values get locked in at some stage in our development, then actions to improve our values now may lead to us being guided by better values for millions of years (see Ord 2020: 153–8 and MacAskill 2022: 78–9). Or more generally, if there are multiple longterm equilibria for our society, early actions may be able to change which one we end up in, and so have lasting effects.

Longterm trajectories

Consider the trajectory of human history up to the present day, and then imagine some of the ways it could continue far beyond, into the distant future. We might have a short future of unconstrained technological progress ending in an extinction event in the next century. Or we might have a much longer future where we live modestly, in harmony with nature, for a typical species lifetime of a million years. Or perhaps we might have a future of epic proportion: spanning trillions of years and billions of worlds.

The purpose of this essay is to find a way of formally representing such trajectories that helps us understand and compare the many ways we might have lasting effects upon them. Doing so involves a balancing act between abstracting away enough details to make the representation workable, while leaving enough to provide the power to make fruitful comparisons.

There are many ways we might represent such trajectories. For example, we could characterise humanity's situation at any one time as a point in a multidimensional space, with dimensions that represent all the key variables for our civilisation, such as our technological capacity, our co-ordination, and our wisdom. We could then imagine the trajectories as the paths our civilisation might trace out in this space as its combination of attributes changes.[2] Such an approach has been suggested by Bostrom (2013) and Beckstead (2013: 69–73).

Or we might focus on a single dimension (such as our technological capacity) and consider a graph of how this could rise or fall over time. Bostrom (2013) and Baum et al (2019) adopt such an approach, using it as a way of illustrating and teasing out different *qualitative* scenarios, such as stagnation, collapse with the possibility of later recovery, or an ephemeral success followed by premature extinction.

I want to explore a somewhat different approach. Like Bostrom, and Baum et al, I will focus on a single representative dimension of humanity and how that evolves over time. But I use a special dimension, chosen to allow quantitative evaluations and comparisons of trajectories.

My primary focus is the value achieved by humanity over time. On the vertical axis is the *instantaneous* value of humanity at any moment in time. This means that the total area under the curve represents the cumulative value of humanity over all time. For example, we could think of the vertical axis in terms of value per year, the horizontal axis in terms of years, and thus the area under the curve in terms of value.

Exactly what this represents depends on our theory of value. As a simple example, if our theory of value were classical utilitarianism, then the height of the curve would be the total amount of happiness minus suffering occurring at that time, and the total area under the curve would be the cumulative amount of happiness minus suffering created in the whole of human history.

But the approach is flexible enough to encompass a wide range of conceptions of value. It could have a richer story of what contributes to an individual's wellbeing; it could ascribe intrinsic disvalue to inequality in wellbeing (at a given time); or it could ascribe value to things other than wellbeing, such as art, knowledge, achievements, or the environment.

The main constraints are:

1. Value is something that is at least roughly quantitative, such that it makes sense to add it up.

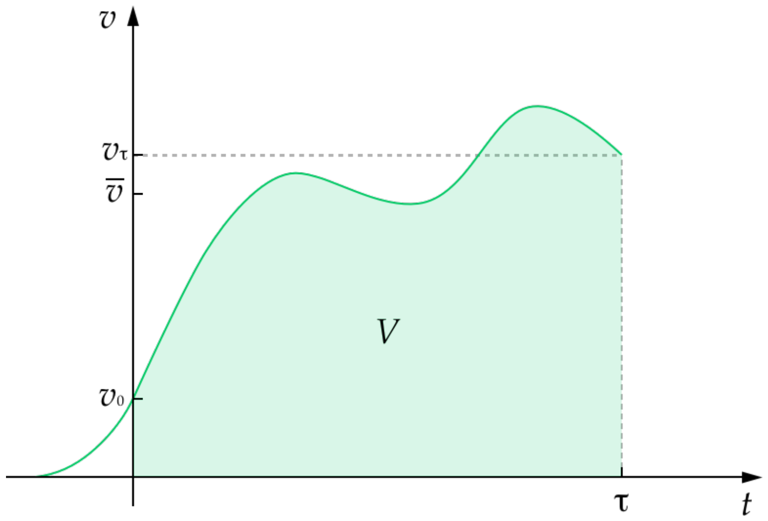
[2] An interesting choice for such an approach is whether to distinguish trajectories that trace the same path through state space but proceed at different speeds or linger at different points on the path.

2. Value is separable across time, such that we can ‘date’ contributions to intrinsic value to a particular time and find the total value by the integral (or sum) across these times.[3]

This is compatible with a range of theories of population ethics including: the total view, the critical level view, a version of the average view that values the integral of the average wellbeing at each time, some person-affecting views, and theories that have diminishing marginal value for additional population at a given time. But it is incompatible with views that are not time-separable (see Broome 2004), such as the average wellbeing of all people who will ever live, or a view with diminishing marginal value on the total number of people who will ever live

The chief reason for adopting this framework is that by focusing on value as the key dimension, it allows us to evaluate and compare different trajectories — and even different *changes* we might make to a trajectory. The hope is that this will inform us about how we should strive to change humanity’s future. But an important limitation is that it will only produce considerations related to value. If there are important parts of morality that are unrelated to value (such as personal freedoms or unbreakable rules) the framework will remain silent about those. It will just tell us how much value is at stake and leave these other considerations up to us.

To see what this framework allows us to do, let’s start with an illustrative trajectory for humanity.



[3] In my presentation of the framework, I’ll treat things as if value is ascribed to particular instants and so is integrated over time. But a lot of what I say would be the same if it were instead ascribed to longer periods and was summed over time. The main issues are that we may no longer be able to apply the concepts of a continuous or smooth curve and that it may be messy if an intervention advances or delays progress by some fraction of a time period. Neither are big issues if the periods were very short (e.g. minutes) but it may become a substantial issue if the relevant periods are generations or centuries.

Figure 1. An illustrative longterm trajectory of humanity. On the vertical axis is the instantaneous value of humanity and on the horizontal axis, time. This means that the shaded area under the curve represents the cumulative value of humanity's future.

This trajectory begins with rapid and escalating growth in the instantaneous value of humanity that slows towards a peak, then gradually declines before rising to an even higher peak. After a final period of decline, humanity's trajectory abruptly ends (perhaps in catastrophe; perhaps after achieving all it could). I have labelled some important features that apply to any trajectory:

t ≡ time (in years)

$t = 0$ ≡ the present time

τ ≡ the endpoint of humanity's future = the duration of humanity's future

v ≡ the instantaneous value of humanity

$v(t)$ ≡ the instantaneous value of humanity at a given time

(the function $v(\cdot)$ itself can be used as a name for the whole trajectory)

v_0 ≡ the instantaneous value of humanity at the present time = $v(0)$

v_τ ≡ the instantaneous value of humanity at its final time = $v(\tau)$

$\overline{v}$ ≡ the average instantaneous value of humanity from time 0 to τ

V ≡ the total value of the future = $V = \int_0^\tau v(t).dt$ = $\overline{v}\,\tau$

As we have seen, the duration of our future could be truly vast. Most species in our position could look forward to about 10,000 more centuries, and with our unique capabilities we have the potential to last for billions of centuries. So one way in which the longterm future could be vastly more valuable than our current century is through its duration, τ. The assumptions embedded in this framework imply that the value of our future scales linearly with this duration — other things being equal, a future of a thousand times the duration is a thousand times better.

A second way our future could be vastly more valuable than the present is by being much more valuable at any given time. This could be true if we are able to build fairer and more just societies with much less of what is bad in life. It could also come from us each having much more of what is good in life. Since our peak experiences far outshine the average, there is room for each of our lives to be vastly better if only we could spend longer at those heights.

And civilisation of the future may also be much more valuable through being so much larger than what we have today. There are more than 100 billion planets in the Milky Way, so the scale of our future civilisation could be vastly greater than it is now (Cassan et al. 2012, Ord 2021). And this may matter a great deal. Or it may not. It isn't clear how to evaluate such increases in the scale of humanity at a time, and there is much disagreement. But it is a plausible way that the value of humanity at future times may be much much greater than it is today.

To see how much more valuable than our own time the entire future could be, let's shade in the value of our current century on the trajectory we considered earlier:

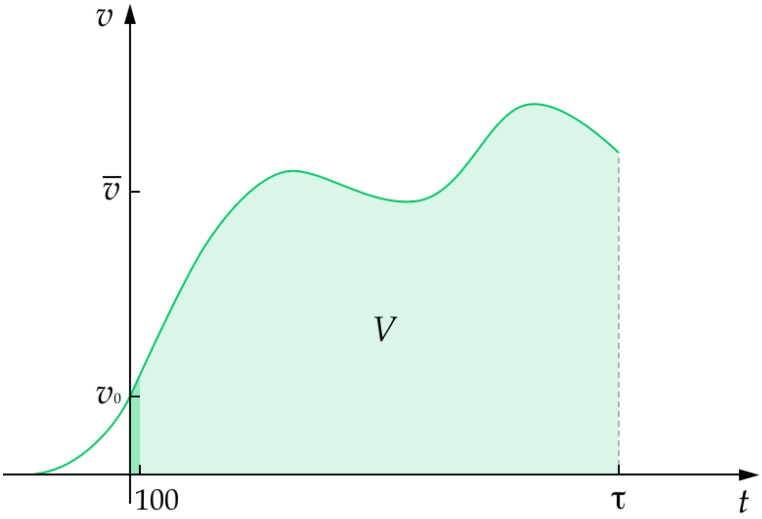


Figure 2. Comparing the value of the next 100 years (shaded dark green) to that of the entire future (shaded light green). Because the length, height, and shape of the future trajectory are extremely uncertain, this diagram shows the key parts that need to be compared but is unlikely to represent them to scale.

Now consider the ratio of the value of our present century to the value of the entire future. The current instantaneous value of humanity is v_0 units of value per year, and so (assuming the average value across the century isn't radically different from the value right now) the value of our century is roughly v_0 times 100. The value of the entire future is V which can also be expressed as $\bar{v}$ τ. So the ratio between these is roughly:

$$v_0\, 100 \;:\; \bar{v}\, \tau$$

We could also think of this in terms of the question:

'How much more valuable is the entire future than the current century?'

Rearranging the expression above, we get a multiplier comprised of two factors:

$$\frac{\bar{v}}{v_0} \cdot \frac{\tau}{100}$$

The first factor is how much better the average century is compared to our own, and the second is how many more centuries there are. These are two quite different ways the future could be much more valuable than the nearer term. And each of these factors might be truly vast — quite possibly a billion or more.

Put another way, the sheer duration of the future is more than enough to get longtermism off the ground. And the sheer scale (or quality) of the future at any one time could be too. Either one alone could suffice. And yet because they are multiplicative, there is also a very real possibility that the true scale of future value could be more than a billion *billion* times that of the next century.

If we were just trying to make the case for a greater focus on the longterm effects of our actions, this observation would be excessive. The most robust and persuasive case for focusing on the long term would be to focus on the most widely accepted part — its duration — and leave the heightened value at a time out of it. But as my aim here is not to argue for longtermism, but to develop a technique for comparing different ways of shaping the longterm future, we need to keep track of both dimensions. Later, we will see that some interventions scale in value with just one or the other of these factors, making comparisons between these interventions depend crucially on the relative sizes of the factors.

Aims

It is worth taking a moment to explain the aims of this framework, and how it fits with work in economics and ethics.

The chief aim is to help us understand and compare different ways of altering the longterm trajectory of humanity. The framework's focus is on the cumulative value of humanity over its entire duration. In particular, on how changes to a range of key parameters of that trajectory affect the total value achieved. While the changes to the parameters may be relatively small (e.g. reaching each point in our development one year earlier, or lowering extinction risk by 1 part in 1,000), their effects will often be vast, as they may be felt over the entire future.

Achieving this aim requires the framework to be quantitative, so that we can use mathematical techniques to analyse what is happening. But many of the results will be qualitative. For example, showing that one kind of intervention scales in value with the duration of our future while another one does not. Understanding the ways that different interventions scale with the shape of the future helps us see when one intervention is really of a different kind to another, and so will help us develop useful categories of longtermist interventions.

I will illustrate many of the ideas with diagrams. These are designed to clearly show what intervention is being considered and how its consequences unfold. Doing so almost *requires* that the diagrams are not drawn to scale. One reason is that we are primarily considering changes that are in some way small compared to what is displayed on the graph — changing things by 1 part in 1,000 wouldn't show up clearly on a scale diagram. But this is a smaller loss than you may think, since there is so much uncertainty in the scale and duration of our future anyway that one really can't hope to draw it to scale in any definitive way. But we do need to bear in mind that in a true-scale diagram, the entire duration of human history to this point in time may be a vanishingly narrow sliver at the start of the trajectory — brief in time compared to the aeons ahead and perhaps meagre in instantaneous value compared to what will be able to be achieved with our full maturity.

One useful point of comparison is with supply and demand curves in economics. These are quantitative curves but are often drawn in a highly stylised manner. They help make qualitative distinctions and develop our intuitive understanding of how various effects scale as a parameter is changed.

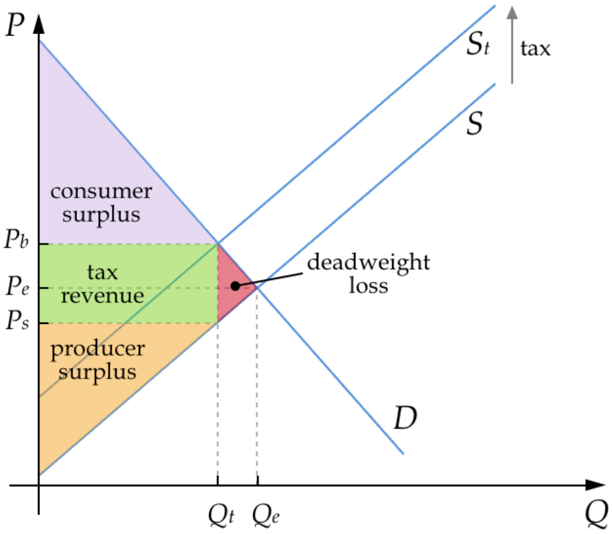


Figure 3. Supply and demand curves showing the effect of adding an excise tax. Like the diagrams I will introduce, this consists of studying what happens to the areas between certain curves as those curves are shifted or adjusted. And its key insights can be gleaned even when the curves are simplified and not drawn to scale.

The theory of longterm trajectories involves aspects of both ethics and economics, and could be studied within either discipline.

All the objects of study — value, time, choices, consequences — are at home within ethics. But the use of mathematical methods to study them is not so common there. Nor is the focus on value over such large scales, or on choices that shape the course of history.

In contrast, the wide scope of optimal planning over the whole future is familiar in economics, as are the mathematical methods used to analyse it (e.g. in the study of optimal growth or intergenerational equity). But the thing being optimised — value — is a little unusual. It is closest to utility, but can have several differences, most notably that it is assumed to be a ratio-scale quantity and to be time-separable. This is what allows us to treat it as a flow of instantaneous value over time and see total value as the integral of that flow (see John Broome (2004) for how this can be done in ways that should satisfy both the philosopher and the economist).

The other main point of departure from most economic analysis, is that I use a zero rate of pure time preference.

Discounting

Economists frequently study flows of benefits and costs over time. They typically evaluate such flows using a technique called discounting. This means applying a mathematical function that reduces ('discounts') the relative importance of costs and benefits that occur further into the future. They do this for a variety of different reasons.

One key reason is that economists are usually considering monetary benefits or costs. Money doesn't translate directly into value, so its value may depend on when it is received or paid. Money often needs to be discounted due to empirical facts about the interest rate, the growth rate, and the diminishing returns to wellbeing from having more money. Earlier monetary costs are magnified compared to later ones, because they mean you miss out on investing the money over the intervening years; later monetary benefits are typically worth less than earlier ones because the people who receive them are richer, so get less value from each dollar. These are good reasons for discounting monetary benefits and costs, but they won't apply here as we are directly concerned with evaluating streams of value (or utility) itself.

Another reason offered for discounting the future is that people have a brute preference for value to come sooner rather than later ('pure time preference'). For example, a person may simply prefer one treat now to two later. It is not actually so clear whether people do have such preferences, since it is difficult to experimentally distinguish between people having a considered preference for immediate benefits versus suffering weakness of will where they act in a short-termist way against their own better judgment.

But regardless of whether people have such preferences with respect to their own lives, experimental evidence suggests they don't have them regarding benefits to other people, such as one's children or future generations. Faced with such choices, people are roughly indifferent to the timing of the benefit or cost (Frederick 2003). Since these are the kinds of choices under discussion when considering the longterm future of humanity, and because there are also strong moral arguments for treating people equally whenever they happen to live, I will follow the consensus of moral philosophers (and some eminent economists such as Pigou 1920:25, Ramsey 1928:543, and Harrod 1948:40) in rejecting this kind of discounting.

A third reason for discounting is based on risk. An individual might rationally discount the value of benefits they would receive in their eighties or nineties on account of the reduced probability that they are alive to receive them. Similarly, the chance that humanity is still in existence will monotonically decline over time.

This could certainly provide reasons to discount future benefits (Ng 2005). But it doesn't apply to the methods used in this chapter. Here I'm focusing on the *ex post* value produced by an intervention in a particular outcome. Risk could then enter the picture by considering prospects over these outcomes. If this includes uncertainty over the length of the trajectory, that will produce an effect similar to discounting, though one that isn't baked in via an exogenous hazard rate (such as Stern 2007). Instead, it would be more flexible and expressive, allowing for variable hazard rates and for the choices we make to affect those rates. But this deeper treatment obviates any need to discount the individual *ex post* outcomes we are considering in this chapter on grounds of risk.

Finally, some economists (e.g. Koopmans 1960) have suggested that whether or not there is a good reason, we simply *must* discount future value because otherwise the sum of value over time becomes infinite, causing intractable mathematical problems when evaluating or comparing different choices.

There are real challenges here, and they affect all approaches unless certain restrictions are made. Even exponential discounting fails to resolve the problem unless there are restrictions to prevent the possibility of instantaneous value growing exponentially over time. In the present work I avoid the problem by restricting the scope to trajectories that are finite in duration.[4] This is not much of a restriction in practice. Instead of considering infinite flows of value, I just consider finite flows with longer and longer durations. The duration of humanity's future appears as a parameter, τ, and we can ask questions such as:

- Which intervention is superior in the limit of large τ?
- Where is the cut-off for τ beyond which intervention 1 has a better effect than intervention 2?
- How does the value of this intervention scale with τ?

This framework thus has an unusually low amount of discounting compared to most related work in economics. And this will cause some challenges.

[4] When extending this work to deal with uncertainty, the challenges would return due to the possibility of infinite expected value even when the duration is certain to be finite (e.g. consider a survival function with a tail that decays as $1/t$). In such cases I'm optimistic about resolving some of the challenges by assigning these hyperreal expected values, in a manner related to Bostrom 2011 and Pivato 2008.

When there is ample discounting of the future, one doesn't have to worry much about how to model the very longterm impacts of the intervention you are studying. The impacts of the longest run effects are so dampened by the discounting that their entire value from 100 years out to eternity is usually bounded and small.

This is convenient. For without discounting, the comparisons between different interventions could be mainly driven by the assumptions about their longterm effects — assumptions that are quite unreliable if they were not the focus of the economic analysis. Heavy use of discounting (especially pure time preference) avoids this.

But it can also throw the baby out with the bathwater. What if the main effects of our everyday policies really are their longterm effects? Or what if we want to study the set of interventions targeted towards changing the long run future? If so, then heavy use of discounting, while convenient, would be inappropriate. These longterm effects are the very object of our study. They need to be illuminated and clarified, rather than swept aside. Indeed, we could even go so far as to say that one of the purposes of this framework is to allow us to analyse and compare very long-reaching effects of our actions without needing to discount them.

Idealised changes to the trajectory

When analysing lasting changes to humanity's longterm trajectory, we are often interested in marginal changes — that is, relatively small adjustments to our trajectory from the present time, $t = 0$, onwards.

Why marginal changes? Since the overall trajectory could be so much longer than everyday timescales and the average instantaneous value so much greater than what we see today, even changes that are very large in today's terms could be relatively small compared to the whole future. Relatively small changes to the longterm future may thus be the best we can do. And yet they may still matter a great deal in absolute terms, as their value could accumulate over deep time. Marginal changes are also easier to analyse, allowing us to imagine everything else being roughly the same despite the change. Of course, marginal changes are not the whole story, but they are a key component and a good place to start.

Mathematically, we will be analysing a marginal change to some key parameter of the trajectory of humanity, adjusting it by some small amount. We will use the symbol δ (as in δt) to refer to these marginal changes in a parameter (reserving the capital letter Δ for large changes, to be studied at a later date). Each marginal change will transform the trajectory $v(\cdot)$ from 0 onwards. We shall represent the transformed trajectory as $v^*(\cdot)$ and its endpoint as τ^*. We can then examine the effects of this change upon the value of the future (the integral of the trajectory from 0 to τ^*).

In general, evaluating the effects of even a small change to the shape of the future trajectory would require detailed knowledge of the shape of the default trajectory — the course history was going to take before we intervened. Given our ignorance of this shape, that would make such evaluations extremely difficult. But there are a family of idealised changes to the trajectory that can be evaluated without detailed knowledge of the default shape. Instead, they depend only on a few of the key parameters we've seen earlier, such as the duration of the future, τ. While these parameters are *also* unknown, the quantities we care about will now be expressible as simple functions of these unknowns.

The kinds of idealised change we will explore are:

- Advancements
- Speed-ups
- Gains
- Enhancements

In reality, the interventions open to us will be much messier than any of these idealised versions. But due to their simplicity and tractability, the idealised changes will provide a useful starting framework for analysis.

Advancements

One way to change the future is to advance progress. If we think the future is likely to be better than the present, we could try to reach those higher levels of instantaneous value sooner.

There are, of course, many kinds of progress: scientific, technological, societal, moral, and more. And each of these has many different strands within it. A nuanced approach to advancing progress therefore involves the possibility that some of these advance more than others, which could have complex effects on the shape of the trajectory.[5]

But it is useful to also ask the simpler question: what if we could shift the trajectory of humanity's instantaneous value earlier by some small amount of time?[6] This may

[5] Carl Sagan (1994: 316–17) suggested that the fundamental challenge of anthropogenic existential risk stems from our technological progress outstripping our progress on becoming wiser as a civilisation. Bostrom (2014: 228–46) has a good account of the importance of advancing some kinds of progress more than others. For much more on the idea that humanity should act to advance progress (in all its nuance), one could look to the burgeoning field of Progress Studies (Collison and Cowen 2019).

[6] Bostrom (2003) explores a similar approach, comparing advancements to existential risk reduction. Cotton-Barratt (2015: 9–12) models it very similarly to me, though applies the model to advancing progress in building a social movement.

correspond roughly to advancing all forms of progress by that same amount of time. I am not meaning to suggest that all attempts to advance progress will behave like this, but that it is a clean transformation of the trajectory which captures part of what advancing progress is about, and which could be helpful in thinking about real attempts to advance progress.

We shall say that an *advancement* is a change to the default trajectory $v(\cdot)$, where the future trajectory is shifted left by some small amount of time δt. More formally:

$$v^*(t) = \begin{cases} v(t) & , t \leq 0 \\ v(t+\delta t) & , t > 0 \end{cases}$$

(Note that this way of modelling advancements involves a discontinuous jump in the trajectory from where it was at 0 to where it would have been at δt. I don't mean to suggest that it really would jump like this — I presume it would instead continuously rise to that level over a similar timescale to δt. But the difference this makes to the final analysis is small and not worth additional complexity.)

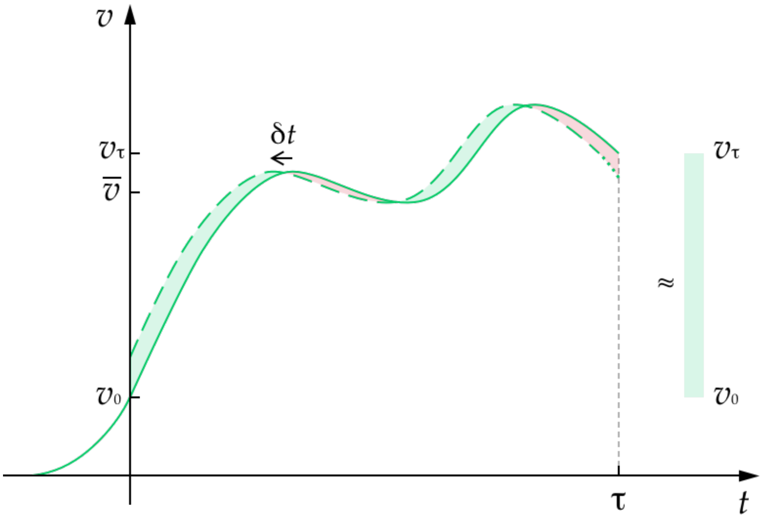


Figure 4. An *advancement*. The dashed trajectory represents a change to the default trajectory where each instantaneous value level is reached δt years earlier. This corresponds to shifting the default future trajectory δt units to the left. (As in all these diagrams, the scale of this shift is exaggerated here for clarity.) The difference in value between these trajectories is equal to the sum of the shaded areas between the curves, where the green areas count positively and the red areas negatively.

This definition so far leaves open the question of whether the end time also gets shifted. If τ is an exogenous end time for humanity (e.g. the sun burning out, or a collision point with a large asteroid) then we might expect it to stay fixed. But if it is endogenous to our activities (e.g. extinction via a dangerous advanced technology), we might expect it to be brought forwards by δt along with everything else. Both cases are plausible, and we'll examine them separately.

Let's begin by considering an advancement with an exogenous end time. This raises a new question of what values $v^*(t)$ takes between $\tau - \delta t$ and τ. I will assume that there is some standard continuation of the default trajectory $v(\cdot)$ beyond τ — what would have happened to humanity by default if we hadn't gone extinct at that time. Technically we can suppose that $v(\cdot)$ is defined and continues beyond τ, it is just that we only show (and integrate) the curve up to τ.

Given this approach, the value of an advancement is equal to the sum of the areas of the green regions in *Figure 4* (where the new trajectory is superior) minus the sum of the areas of the red regions (where the default trajectory was superior). For trajectories with many turning points, these regions can get very complex, and their magnitude unclear. But happily, this difference in values between the old and new trajectories can be approximated with a very simple expression which doesn't depend on the shape of the trajectory.

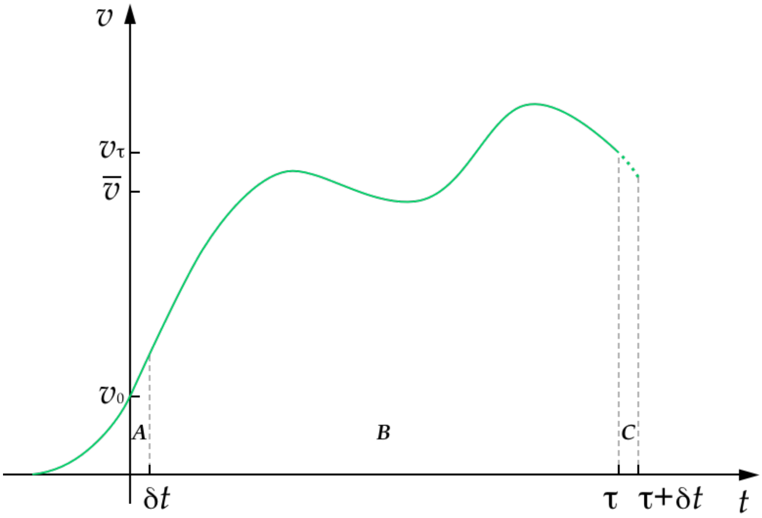


Figure 5. Decomposing the future trajectory into pieces that will help us find the value of an advancement.

Figure 5 shows the shape of the default trajectory, with a dotted green line showing how it would continue beyond τ for the next δt. We can divide the area under this curve into three regions *A*, *B*, *C*. The default trajectory ends at τ. So the value of our default future, *V*, is simply given by:

$$V = A + B$$

In contrast, the alternate trajectory with the advancement skips over piece *A* and consists of pieces *B* and *C*.

$$V^* = B + C$$

So the amount by which V^* is superior to V is:

$$V^* - V = C - A$$

Under certain conditions, we can approximate the sizes of C and A. This depends on how quickly $v(\cdot)$ is changing at 0 and τ. If the instantaneous value is changing relatively slowly at 0 (i.e. if $v_0 \gg |v(\delta t) - v_0|$) then we can approximate A by a narrow rectangle with height v_0 and width δt:

$$A \approx v_0\, \delta t$$

Similarly for C and τ:

$$C \approx v_\tau\, \delta t$$

This gives a simple approximation for how much better V^* is than V (a difference we shall denote δV):

$$\delta V \;\equiv\; V^* - V \;\approx\; (v_\tau - v_0)\, \delta t$$

In many circumstances with a marginal advancement this will be a close approximation. For example, if the instantaneous value of humanity changes by less than 5% each year, then these approximations for A and C will be correct to within 5% for advancements of up to a year. And an advancement of an entire year would be very difficult to achieve: it may require something comparable to the entire effort of all currently existing humans working for a year. Even with a very leveraged opportunity, we might expect the kinds of advancements under consideration to be more like days than years, in which case the approximation should be accurate to better than 1 part in 1,000.

This approximation for δV has no dependence on the shape of $v(\cdot)$. It depends solely on its values at two particular times and the size of the advancement. Indeed, the value of an advancement doesn't even have any dependence on the duration of humanity's future, τ. Looking closer we can see that even the precise values of A and C depend only on the shape of $v(\cdot)$ in the immediate vicinity of 0 and τ, and don't depend on τ.

So despite an advancement being a kind of lasting change to the trajectory of humanity — a change whose effect remains at full strength over our entire future — its value doesn't scale with the length of this future. Whether humanity's future lasts a million years or a billion years doesn't affect the value of an advancement, except in-as-much as we might hope that we reach higher heights the longer we have. But if our world (or our value system) is such that the instantaneous value of humanity plateaus, then duration doesn't matter much to the value of advancements. And if our future trajectory is such that the later instantaneous values aren't much different to those today, then advancements would have very little value at all.

It is interesting to note that the kind of preference motivating advancements — for things to happen sooner rather than later — can arise even from the perfectly patient perspective we are considering in this chapter. It isn't about moving some fixed amount of value earlier in time, but about creating more value. For example, whenever the slope of the trajectory $v(t)$ is increasing over the long run, shifting this trajectory earlier in time makes things better on average across all subsequent times. And if the slope is also bending upwards, the amount by which things are being made better at each time is also increasing over time.

We've focused so far on the case where the end time is exogenous, but what if it is endogenous? The simplest way to model how an advancement changes an endogenous end time is to say that the end time shifts to the left by δt along with everything else (i.e. $\tau^* = \tau - \delta t$).

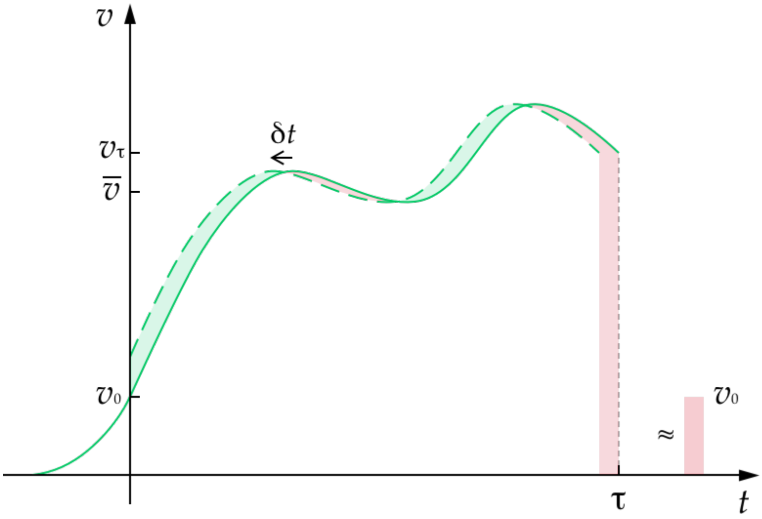


Figure 6. An advancement with an endogenous end time.

If so, the area under the new trajectory (V^*) will still skip over the first piece (A) in Figure 5, but will end before it would get the compensating benefit of the third piece (C). So the value of the future is just that of the middle piece (B):

$$V^* = B$$

So

$$\delta V = V^* - V = B - (A + B) = -A$$

$$\delta V \approx -v_0 \, \delta t$$

This is a substantial difference. Now the only real effect of the advancement is to skip the value of the next δt of humanity's trajectory. So unless humanity currently has a negative instantaneous value, an advancement with this kind of endogenous end time actually makes the future worse. In this case it really is just about moving the future benefits earlier in time rather than increasing the amount of benefits in the future.

Note that this conclusion is sensitive to the precise nature of the question we are asking. We have been comparing two different trajectories — two particular ways the world could unfold. But a natural extension of this framework could consider uncertainty by instead comparing two probability distributions over trajectories or by comparing trajectories with associated hazard curves. This could change the conclusion. For example, if we made the assumption that an advancement by δt skips over the risk in the next δt of our trajectory, this could make the value of the advancement positive again, even when risk of ending the trajectory is endogenous (i.e. when all subsequent risk gets advanced too).[7]

In summary, when the end time is exogenous, it is easy to see how advancing progress across the board could improve our future: roughly speaking, it could replace a period at current instantaneous value with one at the final instantaneous value. This value could scale with the ultimate size of human endeavours in the future, even if it won't scale with the duration of humanity's future. But when the end time is endogenous this kind of justification for advancing progress won't work. On the simplest model, it might just make things worse by skipping over a period at our current level of instantaneous value. So if it is to have good effects, that would need to be due to a more complex story about how it shapes the future. For example, if it skipped over some of the risk from the duration it advanced, or if it advanced progress in some areas more than others, such that it reduced existential risk or changed the trajectory in some way beyond a simple horizontal shift.

Finally, note that the study of advancements also applies to their opposites: *delays*. These correspond to shifting the future trajectory to the right. They fit cleanly into the same mathematical framework by simply allowing δt to be negative.[8] Marginal delays are bad in almost exactly those circumstances when marginal advancements would be good, and vice-versa.

Speed-ups

What if, instead of merely advancing progress, we could permanently speed it up? Perhaps there are ways of organising society that would achieve in 100 years what would have taken us 101, achieve in 1,000 years what would have taken us 1,010, and so on. As this would have a proportionally larger effect further into the future, it wouldn't be shifting the future trajectory to the left, but horizontally compressing it by some factor γ_t:

[7] Aschenbrenner (2020) reaches similar conclusions with his economic model.

[8] There is actually a small additional wrinkle, in that you need to define what $v^*(t)$ is between $v(0)$ and $v(\delta t)$. That is, what small piece of trajectory to insert to join the past trajectory to the delayed future trajectory. The most natural answer is to have it be a flat line at $v^*(t) = v(0)$

$$v^*(t) = \begin{cases} v(t) & ,t \leq 0 \\ v(\gamma_t(t)) & ,t > 0 \end{cases}$$

In this way, γ_t could be thought of as the factor by which progress is sped up and we can call this kind of change a *speed-up*. Marginal speed-ups would correspond to γ_t being slightly greater than 1. For instance, the 1% speed-up described in the previous paragraph would correspond to $\gamma_t = 1.01$.

As with advancements, it matters whether the end time is exogenous or endogenous. Let's first consider the exogenous case, as in *Figure 7*.

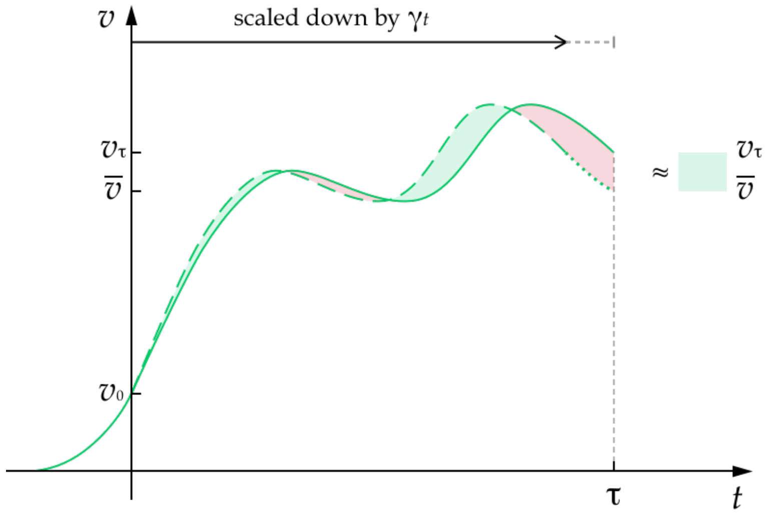


Figure 7. A speed-up with an exogenous end time.

To see how this changes the value of the future, we can decompose the future into two parts. First, note that the average instantaneous value under the dashed green curve is just the same as the average instantaneous value under the default trajectory: $\overline{v}$. And for marginal speed-ups (those where $\gamma_t \approx 1$), the average instantaneous value under the *dotted* green line is approximately v_τ. So the new value of the future, V^*, is a weighted average of these:

$$V^* \approx ((1/\gamma_t)\,\overline{v} + (1 - 1/\gamma_t)\, v_\tau)\, \tau$$

The amount by which it is better than the default is:

$$\delta V = V^* - V \approx (v_\tau - \overline{v})\,(1 - 1/\gamma_t)\, \tau$$

This is a product of three factors: the amount by which the end time is better than the average time, a number just above zero representing the fraction of humanity's duration in the dotted part of the trajectory, and humanity's duration itself.

This whole effect has a positive value if and only if the instantaneous value at the end of the default trajectory is higher than its average instantaneous value.

For speed-ups with endogenous end times (as in *Figure 8*), the duration of humanity is also sped up, with $\tau^* = \tau / \gamma_t$.

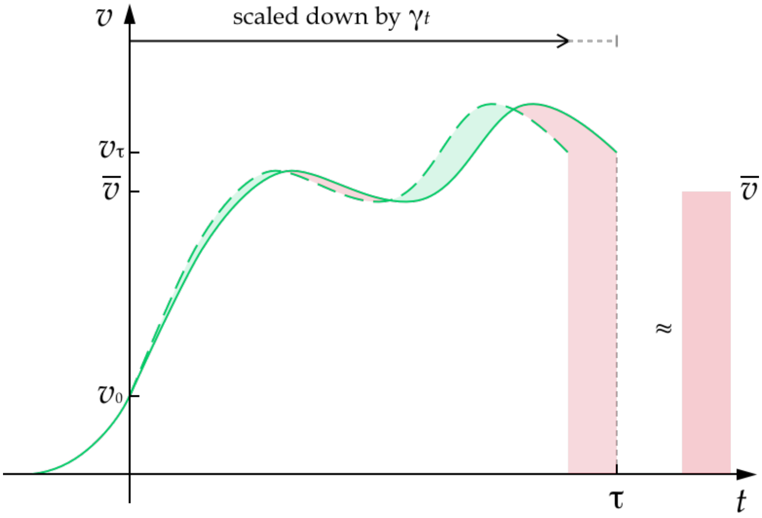


Figure 8. A speed-up with an endogenous end time.

The area under this new (dashed) trajectory is just V/γ_t — a compressed version of the default value of the future. And the improvement it makes on the future is:

$$\delta V = V^* - V = -\overline{v}\,(1 - 1/\gamma_t)\,\tau$$

This is of positive value if and only if the total value of the default trajectory is negative. That's because it corresponds to having all the same levels of value in the future, just spending less time at each one — making the value of the future a little smaller. So again, the distinction between whether the end is exogenous or endogenous is very important for evaluating this intervention. Speed-ups with endogenous end times aren't generically a good thing — they would need to be differentially speeding up progress or reducing risk in some way to be valuable.

How plausible are speed-ups? The broad course of human history suggests that speed-ups are possible. For example, the agricultural and industrial revolutions each seemed to substantially speed up the clock of human progress compared to what came before. And presumably they also sped it up compared to what would have happened if they had never occurred. Of course, they also had other effects, speeding up some processes more than others and causing other changes to the trajectory. But the idealised idea of a speed-up would still seem to capture a key part of what happened. Both revolutions would count as non-marginal speed-ups. More incremental speed-ups may also exist, though we wouldn't expect to be able to identify a marginal speed-up (say 1%) in the noisy historical record.

What would it look like if there were many marginal speed-ups? If there were a succession of them, happening roughly uniformly in time, the overall effect could be exponential. Imagine starting with an upwards diagonal default trajectory then at time 1 introducing a 1% speed-up, at time 2 introducing a further 1% speed-up and so forth. After the first speed-up, the slope is 1% higher than before, after the second it is 1% higher than that (compounding). Since the slope is increasing exponentially with time, so is the curve itself, and thus so is the integral.[9]

It is plausible that some component of the exponential growth humanity has achieved across various dimensions is due to the accumulation of speed-ups like this. But it is not clear whether the key dimension — the instantaneous value of humanity — has actually grown exponentially. It probably has according to theories of population ethics where the instantaneous value is proportional to the population at that time (all things being equal), but not according to most other theories. So this case for successive small speed-ups is more suggestive than proven.

One big challenge for the idea that speed-ups are something for altruists to aim towards is that if a speed-up is possible at all, it seems likely to be overdetermined that it will happen. That is, one could calculate the difference in value between a trajectory where it doesn't happen and one where it does, leading to a high valuation — but that might not be the relevant comparison. If it is overdetermined to happen, then by making it happen now, we are just bringing forward the time when it happens. And if so, then we just have an advancement rather than a speedup. I think this is probably the case for the agricultural revolution (which was so overdetermined it independently occurred in at least five different parts of the world (Christian 2004: 248–87)), though there is more scholarly debate about whether the industrial revolution would have ever happened had in not started in the way it did. And there are other smaller breakthroughs, such as the phonetic alphabet, that only occurred once and whose main effect may have been to speed up progress. So contingent speed-ups may be possible.

The opposite of a speed-up is a *slow-down*. It is what you get when γ_t is less than 1 and corresponds to horizontally stretching the future trajectory. Note that like speed-ups, slow-downs are defined compared to what would have happened by default. So a slow-down could take the form of an event that actively slows down future progress or it could take the form of a choice *not* to pursue some new development we were headed towards that would have sped things up. For marginal changes, slow-downs are good when speed-ups are bad, and vice versa.

Gains

What if, instead of adjusting the timings of the future trajectory, we could directly adjust the instantaneous value itself? An advancement is a shift of the trajectory to the left (which sometimes leads to it rising on average), but what if we could directly shift the trajectory upwards? We can all such a change a *gain*. Mathematically:

[9] If each speed-up also sped up the rate at which future speed-ups would happen, the growth would be even faster, approaching a vertical asymptote at a finite time.

$$v^*(t) = \begin{cases} v(t) & ,t \leq 0 \\ v(t) + \delta v & ,t > 0 \end{cases}$$

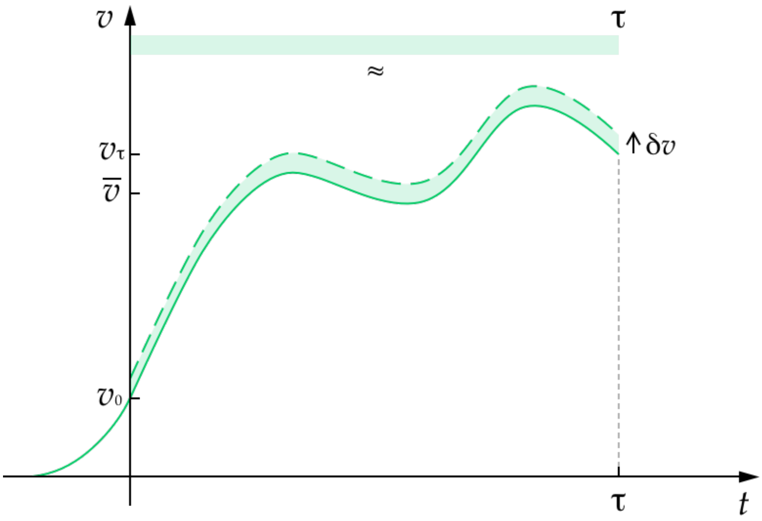


Figure 9. A gain of size δv.

The value of a gain is easy to calculate. The new value is δv units higher at all times, so:

$$V^* = (\overline{v} + \delta v)\tau$$

and

$$\delta V = \delta v\ \tau$$

While the idea of a gain is simple — a permanent improvement in instantaneous value of a fixed size — it is not so clear how common they are. Many kinds of permanent improvements in humanity's wellbeing might be expected to scale in value with other quality of life improvements or with our population. If so, they are likely to be proportional increases in humanity's value at any time, rather than fixed increases, so would not count as gains (they will be addressed next). However, certain limiting cases may count as gains. For example, a fixed improvement in everyone's quality of life in a future where the population doesn't change much would count, as would a fixed improvement in everyone on Earth's quality of life even if additional people came to exist elsewhere, or a fixed improvement in everyone's quality of life according to a system of population ethics where instantaneous value didn't scale with the size of the population at that time.

Such a fixed improvement to each individual's instantaneous wellbeing could come via a direct effect on wellbeing of fixed size, or via a proportional improvement in some instrumental good (such as income) whose effect on wellbeing is logarithmic.

Other examples could be found beyond human wellbeing. For example, a permanent improvement to the wellbeing of animals on earth would behave like a gain (though it would require an adjustment to what $v(\cdot)$ is supposed to be representing). Or consider an improvement to a non-welfarist good, such as saving an ecosystem, a species, or a work of art. For value systems where the value of such things is proportional to how long they last, these would count as gains.

As with speed-ups, a putative gain often faces a challenge regarding whether it was truly contingent, or would have simply happened later. If the lasting benefit would have been achieved later the default trajectory, then it is only temporary so not a true gain. For example, making a scientific discovery may make things better for all subsequent times, but if the default is (as usual) that someone else would have discovered it sometime later, then that improvement is not permanent, so not a gain. If its value lies not just in improving the value at each time, but in allowing us to get to future points in our development sooner, then it may be an advancement instead. One way that something can resist this challenge is if it takes the form of saving something irreplaceable from permanent destruction.

The opposite of a gain is a *loss*. It is what you get when δv is negative, and corresponds to shifting the future trajectory down. Losses could result from adding things of negative value, removing things of positive value, preventing things of positive value being created, or preventing the only chance that something of negative value could be remedied.

Enhancements

It may also be possible to permanently improve humanity's instantaneous value by a given proportion. For example, if we could make every moment across humanity's future 1% more valuable. We can call such an idealised change an *enhancement*, and model it as:

$$v^*(t) = \begin{cases} v(t) & , t \leq 0 \\ \gamma_v v(t) & , t > 0 \end{cases}$$

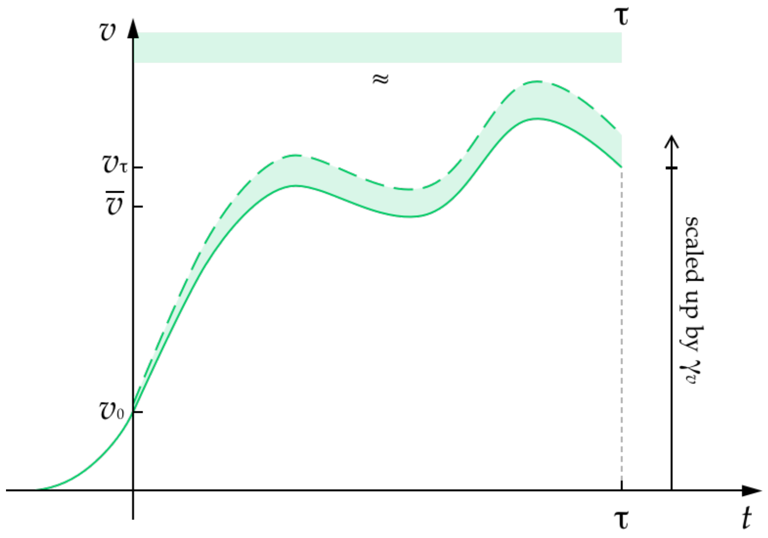


Figure 10. An enhancement by a factor of γ_v.

As this is simply a scaled-up version of the default trajectory, the value is trivial to calculate:

$$V^* = \gamma_v \bar{v} \tau$$

and

$$\delta V = (\gamma_v - 1) \bar{v} \tau$$

There are many kinds of changes to our future that could be modelled as enhancements. For example, improving the quality of life for everyone who will live by a modest proportion (perhaps via improvements to health, technology, prosperity, our lived environment, or our social structures). Alternatively, proportionally increasing the number of people alive at each time would be an enhancement according to certain theories of population ethics. Or improvements in our moral understanding or moral motivation could mean that we produce systematically better outcomes.

As with many of these idealised changes, they face the challenge of why this wouldn't happen eventually, even without the current effort. I think this is a serious challenge for many proposed enhancements. Those that can best resist it may be cases where there is going to be a kind of lock-in (Ord 2022: 153–8). For example, if the values that guide humanity become locked-in at some point in time, then improvements to those values prior to that point could have truly lasting impact.

The opposite of an enhancement is a *diminution*. It is what you get when $\gamma_v < 1$ and corresponds to vertically compressing the future trajectory.

Combinations and variations

Interventions could also produce a combination of these idealised changes. And it is easy enough to allow this in the mathematics. We simply need to keep track of all the locations where a delta or gamma could lurk and allow more than one to take a non-trivial value. The general form is:

$$v^*(t) = \gamma_v\, v(\gamma_t\, t + \delta t) + \delta v$$

The deltas can move the trajectory in any combination of up, down, left, and right. And the gammas can stretch or compress it horizontally or vertically. This allows substantial flexibility in how the trajectory is transformed. That said, the flexibility comes at a cost of increased complexity and reduced clarity about the way in which the intervention is affecting the future and why. My guess is that the individual transformations are the more useful tool.

One form of idealisation has been that these changes to the trajectory will be permanent. But of course, many changes are not. A very general way to represent them is to have a decay curve *d*(*t*), that fades the effect out over time. *d*(*t*) would start at 1 and monotonically decay according to some desired schedule. This could fully remove the effect by some specified time, have the effect asymptote towards zero, or have it decline to some smaller but still positive size. As an example, a temporary advancement could have the equation:

$$v^*(t) = \begin{cases} v(t) & ,t \leq 0 \\ v(t + d(t)\delta t) & ,t > 0 \end{cases}$$

Alternatively, much of the benefit of modelling temporary changes might be gained through the much simpler system of assuming the effect is at full force for some specified duration before completely vanishing. That is less realistic and less flexible but may give most of the benefit with just a single parameter.

Consider a temporary advancement where the effect ends at time, t_e, after which the default trajectory resumes:[10]

$$v^*(t) = \begin{cases} v(t) & ,t \leq 0 \\ v(t + \delta t) & ,0 < t \leq t_e \\ v(t_e + \delta t) & ,t_e < t \leq t_e + \delta t \\ v(t) & ,t_e + \delta t < t \end{cases}$$

For a temporary advancement:

[10] Advancements require this third clause to fill in part of the trajectory that would otherwise be missing.

$$\delta V \approx (v_e - v_0)\ \delta t$$

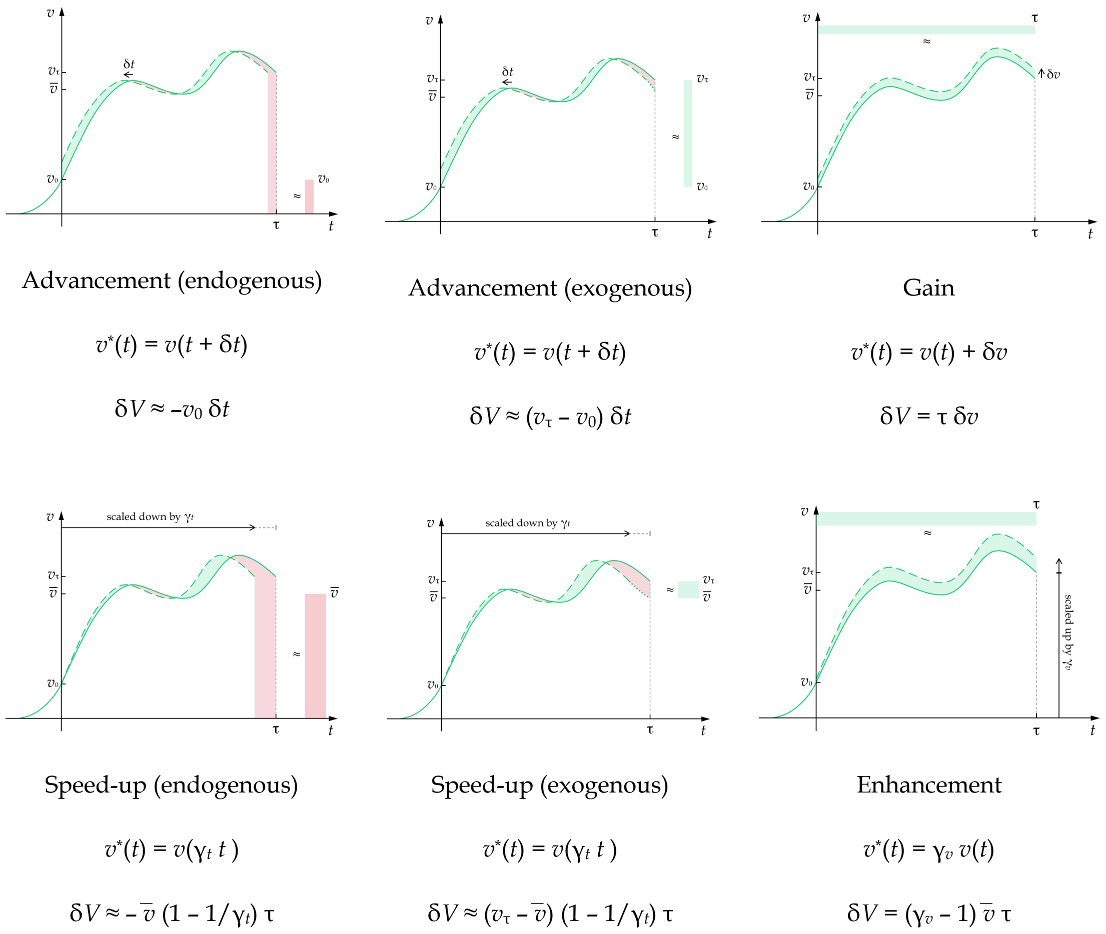
So an advancement by δt would produce the maximum benefit if it persisted until the moment of humanity's peak instantaneous value, but no further. For example, an intervention which advances progress for the entire period in which humanity is ramping up to its full and final scale in the universe, but which does nothing to change the schedule of the long denouement as the stars wind down.[11] The fact that such a temporary advancement would be superior to a permanent one may also contribute to making advancements more likely to be temporary. And while temporary advancements require an extra parameter to specify, they are simpler in some other ways: the distinction between whether the end time is exogenous or endogenous becomes moot, as does the sensitivity to the very final instantaneous value.

We could call temporary changes whose effect lasts for a time on the same scale as τ, *lasting* changes to humanity's trajectory. While any precise cut-off is arbitrary, we might operationalise this as a duration at least one tenth of the entire future of humanity.

[11] Nick Bostrom (2003) was focused on an advancement of this kind.

Comparisons

How do these different ways of shaping the longterm future compare?



Advancement (endogenous)

$v^*(t) = v(t + \delta t)$

$\delta V \approx -v_0\, \delta t$

Advancement (exogenous)

$v^*(t) = v(t + \delta t)$

$\delta V \approx (v_\tau - v_0)\, \delta t$

Gain

$v^*(t) = v(t) + \delta v$

$\delta V = \tau\, \delta v$

Speed-up (endogenous)

$v^*(t) = v(\gamma_t\, t\,)$

$\delta V \approx -\,\overline{v}\,(1 - 1/\gamma_t)\, \tau$

Speed-up (exogenous)

$v^*(t) = v(\gamma_t\, t\,)$

$\delta V \approx (v_\tau - \overline{v})\,(1 - 1/\gamma_t)\, \tau$

Enhancement

$v^*(t) = \gamma_v\, v(t)$

$\delta V = (\gamma_v - 1)\, \overline{v}\, \tau$

Figure 11. A summary of the different idealised changes to humanity's longterm trajectory.

In the right circumstances, any of these idealised changes can be preferrable to any other. It all depends on the sizes of the changes (the deltas and gammas) and the key features of the trajectory (v_0, v_τ, $\overline{v}$, and τ). But there are some important patterns that help us think about what those circumstances are and how plausible they may be.

A useful way to categorise the idealised changes is by whether their impact scales with the heights of instantaneous value we may reach (via v_τ and $\overline{v}$), the aeons we may last (τ), or both. We can see that:

- Advancements scale with $v_\tau - v_0$
- Speed-ups scale with $(v_\tau - \overline{v})\tau$
- Gains scale with τ
- Enhancements scale with $\overline{v}\tau$

It follows that if we consider longer and longer durations for our future, the interventions that don't scale with duration (advancements) become comparatively less important. And similarly, if we consider cases where we reach higher and higher instantaneous values, then gains become less important. Or if we start to acquire a stronger belief that the instantaneous value of the future won't be much larger than that of today, then everything but gains are lowered in importance (and advancements become especially unimportant).

Let's see how the expressions for δV can be used to better understand comparisons between these idealised changes. For example, when is an advancement better than a gain?

$$\begin{aligned}(v_\tau - v_0)\delta t &> \tau\delta v \\ \frac{\delta t}{\tau} &> \frac{\delta v}{v_\tau - v_0}\end{aligned}$$

So an advancement is better than a gain when the proportion of humanity's lifespan that is advanced is greater than the proportion of the gap between current value and final value that is gained. While the units of years advanced and instantaneous value gained didn't directly permit comparison, we see that using this framework the question ultimately comes down to the percentage change in each quantity, which is unitless and therefore comparable.

Moreover, the ratio by which an intervention is superior is just the ratio of these percentages. So advancing progress by a millionth of humanity's lifespan is ten times as good as permanently gaining a ten-millionth of the gap in instantaneous value between now and the end of humanity. And this perspective generalises: the best way of comparing longterm effects often comes down to a comparison of the percentage improvements.

When is an enhancement better than a gain?

$$\begin{aligned}(\gamma_v - 1)\bar{v}\tau &> \tau\delta v \\ (\gamma_v - 1) &> \frac{\delta v}{\bar{v}}\end{aligned}$$

Here one of the biggest unknowns in the study of the longterm future — the ultimate duration of humanity's future, τ — appeared in both expressions and could simply be cancelled out. So the question of enhancements versus gains is not sensitive to this key unknown. We are again left with two percentages to compare: the percentage by which the future is being enhanced versus the percentage of the average instantaneous value that is being gained.

Some of the most important comparisons address how these idealised changes compare to reducing existential risk.

While a full accounting for existential risk unfolding over time would require us to extend the theory to deal with uncertainty (over trajectories or over τ), we can get a lot of value from a highly simplified account. We can consider that there is some probability of existential risk occurring in our time, and model this with the risk occurring right at t_0.[12] We suppose that if this existential catastrophe were to happen, then the value of the future would be extremely small compared to V, and approximate it as zero.[13] For our present purposes of comparing longterm interventions, it doesn't matter how much risk is occurring as it shrinks the value of all the idealised interventions equally.

We can then consider an intervention that increases our probability of surviving this near-term existential risk by a factor, γ_p. For instance, if there were 20 percentage points of near-term existential risk (so an 80 percent chance of survival), and the intervention increased that survival chance by 1 percentage point, then $\gamma_p = 81/80 = 1.0125$. On this model:

$$\delta V = (\gamma_p - 1)\ \bar{v}\ \tau$$

Reducing near-term existential risk is thus another kind of intervention that scales with both $\bar{v}$ and τ. Indeed, the effect it has on the (expected) value of the future is almost identical to that of an enhancement: they both multiply the entire value of the future by some factor. So if prioritising between an enhancement and existential risk reduction it all comes down to which one has the higher factor.

Though remember that for something to be a genuine enhancement, it needs to be truly contingent — a kind of permanent proportional improvement that would never have happened otherwise. Part of the reason reducing existential risk is so important is that it is much easier to meet this bar for contingency — that the loss is irreversible (or nearly so) is built into the definition of existential risk.

We can also compare existential risk to other idealised changes. When would an advancement (with an exogenous end time) be better than lowering existential risk?

$$\begin{aligned}(v_\tau - v_0)\delta t &> (\gamma_p - 1)\bar{v}\tau \\ \frac{\delta t}{\tau} &> (\gamma_p - 1) \cdot \frac{\bar{v}}{v_\tau - v_0}\end{aligned}$$

[12] So long as the timeframe for 'our time' is less than one percent of τ, this way of modelling near-term risk will be fairly accurate.

[13] This is less problematic than it sounds. If the remaining value were, say, a tenth that of V (about the highest remaining value that could still count as an existential catastrophe), then the approximation as zero value still wouldn't change the results much. It would just mean that the true δV for existential risk reduction was 10% smaller than the approximation.

This is a bit harder to interpret, but not impossible. In some situations, we might expect $\overline{v}$ and $(v_\tau - v_0)$ to be of similar magnitude (e.g. if humanity's instantaneous value spends a long time at a high plateau). In that case, advancements are better roughly when $\delta t/\tau > (\gamma_p - 1)$ — when the percentage of the duration of humanity's future that we advance is greater than the percentage by which our survival probability is increased. This makes it difficult for advancements to beat existential risk reduction in scenarios where humanity would have a long lifespan if only it could survive the near-term risks. For example, on a million-year lifespan (that of a typical species) a one-year advancement would be roughly as important as a one-in-a-million improvement in nearterm survival probability — but the latter seems much more achievable.

Even if the simplifying assumption that $\overline{v}$ is similar in scale to $(v_\tau - v_0)$ were not true, it is still very difficult for advancements to beat existential risk reduction, as in order to compensate, this ratio would need to be extreme. It would require a trajectory where the final value was far higher than the average, or a temporary advance up to an intermediate time at the top of a very high and narrow peak in instantaneous value.

It is important to remember that all these equations and comparisons are just for the pure, idealised, changes. Real attempts to improve progress would undoubtedly shift some areas more than others, leading to more complicated effects on the shape of our future. But one upshot of this analysis is to find the situations in which these more complex effects would be necessary.

It seems to me that for attempts to advance progress to be more valuable over the long term than attempts to reduce existential risk, such complex effects would be required. For example, advancing defensive technologies or collective wisdom may reduce existential risk, and advancing moral progress may lead to better values eventually becoming locked in and guiding the future. Even if these kinds of effects might initially seem to be second-order, the fact that they can also scale with τ suggests that they could dominate the longterm value of attempts to advance progress.

Conclusions

In this essay, we have explored a quantitative framework for modelling the longterm trajectory of humanity. By tracking the instantaneous value over time, we've enabled quantitative evaluations and comparisons of different trajectories in terms of the area under the curves. The analysis revealed four different kinds of idealised change to the trajectory corresponding to vertical and horizontal shifts and stretches. Even relatively small changes of these kinds might have vast effects on the value humanity achieves over all time. And given some approximations, these can be usefully analysed and compared. In particular, we've seen that it can be difficult for a pure advancement or gain to rival those interventions like speed-ups, enhancements, and existential risk reduction that scale with both the instantaneous value of the longterm future and its duration. And we've seen that the value of both advancements and speed-ups critically depend on whether they shift the end time as well.

The primary aim has been to help develop a theoretical underpinning for understanding longtermist interventions, but I hope that it will also be of some practical use in finding and comparing tractable interventions of these kinds. And it may also help us weigh the longterm effects of everyday actions taken by people and governments. While such actions are usually aimed at short term effects, given the amounts of value at stake over the long term it is possible that their ultimate impacts are often driven by their longterm effects. If so, this framework could help us better understand their impacts.

All of this was done without discounting. This was made possible by parameterising the lifespan of humanity (which can be treated as exogenous or endogenous) and showing how the value of different interventions scales as a function of this parameter. It also involves explicit modelling of the longterm impacts of our actions, since (unlike in traditional economic analyses) these impacts no longer vanish. While the framework is fully general in terms of the shape of the trajectory, we saw that the values of these idealised changes were remarkably independent of the details of this shape and depended only on a small number of key parameters, greatly helping simplify the analysis.

The framework has substantial room for further development. An important extension is to add uncertainty, either by the fully general approach of comparing probability distributions over trajectories, or by associating a hazard curve with each trajectory. Another extension would be to consider other kinds of changes to the trajectory, including non-marginal changes.

And one could also consider a set of idealised shapes that the future trajectory might take. For instance, what more could we say if we knew the shape was a rapid rise to a long plateau? Or what if there was an exponential rise in instantaneous value all the way to the end time? Such an exponential trajectory has the special feature that an advancement would act just like an enhancement — making every future moment proportionally better. However, despite the recent centuries of exponential *economic* growth, there is reason to be doubtful of exponential growth of intrinsic value, especially over such very long timescales (Ng 1991). Consideration of physical limits to growth might instead suggest an idealised trajectory that grows as a cubic — representing the longterm growth in humanity's resources were we to spread out through the universe.[14] Collating a set of such idealised shapes for trajectories and

[14] Note that the choice of cubic growth makes a tacit assumption that each new location provides its own stream of value (perhaps as a location for our descendants to live, or due to the energy flow of starlight at that location). But it is also possible that the contribution of

exploring how they differ may help us better understand how longtermist priorities depend on the broad shape of the future.

Bibliography

Adams, F. C., and Laughlin, G. (1997). 'A Dying Universe: The Long- Term Fate and Evolution of Astrophysical Objects', *Reviews of Modern Physics*, 69(2), 337–72.

Adams, F. and Laughlin, G. (1999), *The Five Ages of the Universe*, Free Press.

Aschenbrenner, L. (2020), 'Existential Risk and Growth', GPI Working Paper No . 6-2020, Global Priorities Institute.

Barnosky, A. D., et al. (2011). 'Has the Earth's Sixth Mass Extinction Already Arrived?', *Nature*, 471(7336), 51–7.

Baum et al. (2019), 'Long-term trajectories of human civilisation', *Foresight* 21(1):53-83.

Beckstead, N. (2013), *On the Overwhelming Importance of Shaping the Far Future*, [PhD Thesis]. Department of Philosophy, Rutgers University. (Especially pp. 69–73.)

Bostrom, N. (2003), 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', *Utilitas*, 15(3), 308–14.

Bostrom, N. (2011), 'Infinite Ethics', *Analysis and Metaphysics*, 10, 9–59.

Bostrom, N. (2013), 'Existential Risk Prevention as Global Priority', *Global Policy*, 4(1), 15–31.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press.

Broome, J. (2004), *Weighing Lives*, Oxford University Press.

Cassan, A. et al. (2012). 'One or more bound planets per Milky Way star from microlensing observations', *Nature*, 481 (7380): 167–169.

new locations is better thought of as a fixed set of resources (e.g. the materials or energy at that location). If so, it might be better to think of the longterm growth as quadratic. It has also been suggested (Sandberg et al 2016, Ord 2021) that the ultimate physical limits may be set by a civilisation that expands to secure resources but doesn't use them to create value until much later on, when the energy can be used more efficiently. If so, one could tweak the framework to model this not as a flow of intrinsic value over time, but a flow of new resources which can eventually be used to create value.

Christian, D. (2004), *Maps of Time: An Introduction to Big History.* Berkeley: University of California Press.

Collision, P. and Cowen, T. (2019), 'We Need a New Science of Progress', *The Atlantic*.

Galway-Witham, J., and Stringer, C. (2018), 'How Did Homo sapiens Evolve?', *Science*, 360(6395), 1,296–8.

Greaves, H. and MacAskill W. (2021), 'The case for strong longtermism', GPI Working Paper No . 5-2021, Global Priorities Institute.

Ng, Y-K. (2005), 'Intergenerational impartiality: replacing discounting by probability weighting', *Journal of Agricultural and Environmental Ethics* 18: 237–257.

Cotton-Barratt, O. (2015), 'How valuable is movement growth?', Working paper, Centre for Effective Altruism. (especially pp. 9–12).

Ord, T. (2020), *The Precipice: Existential Risk and the Future of Humanity*, London, UK: Bloomsbury.

Ord, T. (2021), 'The Edges of Our Universe', arXiv:2104.01191.

Pivato, M. (2008), 'Sustainable preferences via nondiscounted, hyperreal intergenerational welfare functions', MPRA Paper No. 7461.

Sagan, C. (1994), *Pale Blue Dot: A Vision of the Human Future in Space*. Random House .

Sandberg, A., Armstrong, S., and Cirkovic, M. (2016) 'That Is Not Dead Which Eternal Lie: The Aestivation Hypothesis for Resolving Fermi's Paradox', *Journal of the British Interplanetary Society*, 69, 406-415.

Stern, N H. (2007), *The Economics of Climate Change: The Stern Review*, Cambridge, UK: Cambridge University Press.